DOCUMENT RESUME

ED 400 340 TM 026 075

AUTHOR	Hoachlander, Gary; And Others
TITLE	From Data to Information: New Directions for the National Center for Education Statistics. Conference Proceedings (November 1995).
INSTITUTION	MPR Associates, Berkeley, CA.; National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-96-901
PUB DATE	Aug 96
NOTE	472p.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC19 Plus Postage.
DESCRIPTORS	Data Analysis; *Data Collection; Educational Change; Educational Policy; *Educational Research; Elementary Secondary Education; Futures (of Society); Higher Education; *Information Dissemination; *Outcomes of Education; Research Design; *Research Methodology; School Statistics; Staff Development; Teacher Education; Training
IDENTIFIERS	*National Center for Education Statistics; Opportunity to Learn

ABSTRACT

At the Futures Conference held by the National Center for Education Statistics (NCES) in 1995, discussants from inside and outside the NCES considered the commissioned papers and contributed their expertise. This volume assembles the papers and commentary and summarizes some considerations for policy, research, and practice in future operations of the NCES. The following papers are included: (1) "From Data to Information: New Directions for the National Center for Education Statistics" (Gary Hoachlander); (2) "Tracking Education Reform: What Type of National Data Should Be Collected Through 2010?" (John F. Jennings and Diane Stark); (3) "Where Are We Going? Policy Implications for Data Collection Through 2010" (Christopher Cross and Amy Rukea Stempel); (4) "Enhancing Opportunity To Learn Measures in NCES Data" (Dominic Brewer and Cathleen Stasz); (5) "Teacher Education, Training, and Staff Development: Implications for National Surveys" (David R. Mandel); (6) "'So What?' The Implications of New Analytic Methods for Designing NCES Surveys" (Robert F. Boruch and George Terhanian); (7) "Incorporating Experimental Designs into New NCES Data Collection Methodologies" (Charles E. Metcalf); (8) "Tracking the Costs and Benefits of Postsecondary Education: Implications for National Surveys" (Michael S. McPherson and Morton O. Schapiro); (9) "Special Issues in Postsecondary Education and Lifelong Learning" (David W. Breneman and Frederick J. Galloway); (10) "Large-Scale Video Surveys for the Study of Classroom Processes" (James W. Stigler); (11) "Education and Work: Curriculum, Performance, and Job-Related Outcomes: (Peter Cappelli); (11) "Administrative Record Opportunities in Education Survey Research" (Fritz Scheuren); (12) "New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education" (Glynn D. Ligon). Appendixes describe the contributors and discuss the agenda of future NCES conferences. References follow each chapter. (Contains six tables, four exhibits, and six figures.) (SLD)

TM

ED 400 340

NATIONAL CENTER FOR EDUCATION STATISTICS

Conference Proceedings

FROM DATA TO INFORMATION

New Directions for the National Center for Education Statistics

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM026075

U.S. Department of Education
Office of Educational Research and Improvement
NCES 96-901

BEST COPY AVAILABLE

NATIONAL CENTER FOR EDUCATION STATISTICS

Conference Proceedings

FROM DATA TO INFORMATION

New Directions for the National Center for Education Statistics

Gary Hoachlander
MPR Associates, Inc.

Jeanne E. Griffith
John H. Ralph
National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement

NCES 96-901

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208–5574

August 1996

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *From Data to Information: New Directions for the National Center for Education Statistics*, NCES 96-901, by Gary Hoachlander, Jeanne E. Griffith, and John H. Ralph. Edith McArthur, project officer. Washington, DC: 1996.

Contact:
Project Officer
Edith McArthur
(202) 219–1442
FAX: (202) 219-1575

Commissioner's Statement

In the fall of 1995, the National Center for Education Statistics (NCES) held a conference to stimulate dialogue about future developments in the fields of education, statistical methodology, and technology, as well as to explore the implications of such developments for the nation's education statistics program. This "Futures Conference" was unique for NCES because it attempted to combine considerations in all of these fields in order to stimulate the cross-fertilization and generation of ideas that might not emerge when discussing the topics separately. At this conference, the authors presented commissioned papers on targeted issues that were expected to be important over the next few years, and the discussants provided their comments.

From several perspectives, I believe the conference was highly successful. First, staff from NCES actively participated in all of the deliberations. As a result, they became personally engaged in the *process* of considering alternative futures for their agency. Since the "corporate culture" of this agency is to solicit and build on staff creativity, their participation and interest in this conference was vital. Second, both the formal and informal discussions generated many new ideas. The conference, as such, accomplished far more than the collection of commissioned papers alone could have because of the active interplay of ideas. Finally, many stakeholders in NCES's future saw this conference as a clear signal of the agency's commitment to continued improvement of the usefulness and quality of our surveys and data products. The stakeholders' positive response to the meeting was further reinforced by their expressions of interest in continuing to help in important ways. The success of the conference lies not in the sum of the individual presentations, rather in an overall perspective that provides guidance toward the future.

This publication will serve as a concrete reference to ensure that the stimulating ideas exchanged at the Futures Conference are not forgotten. While the quality of the discussion at the meeting was exceptional, one cannot expect to absorb everything said during a two-day conference. Thus, it is important to have a record that the participants can refer to this year, next year, or five years from now. Moreover, this publication will provide a way to share those ideas with others who could not participate in the conference. For instance, NCES has many customers and other stakeholders who have expressed keen interest in the conference proceedings and whose advice and considerations are welcome as a means to sustain the dialogue about NCES's future.

It is clear that if NCES wants to continue as a key player in providing information for education policy and decision making to the American public, policymakers, education researchers, and educators nationwide, it must continually reevaluate its program and products. In the future, we expect that NCES will receive requests for more of the kinds of products and services that it already provides. Also, we expect demands for new perspectives—on covering new topical areas, implementing new technologies, and adopting new methodologies. Already, major recent changes in the field of education are shaping our future program—for example, widespread innovations to achieve education reform, efforts to adopt both curriculum and performance standards, and examination of education in the United States within an international context. Not only are methodological advances creating opportunities to produce statistics in ways that may be more efficient and effective, but also technological developments are changing the world in which we create data and disseminate our products more rapidly than ever before. The Futures Conference and this publication provide a new vision for NCES—a vision that acknowledges the constraints on the resources of governmental agencies at the end of the 20th century, as well as clearly emphasizes the opportunities that can be achieved with innovative methodologies and technologies and through close attention to the priorities for statistical knowledge in the field of education.

This contribution to envisioning NCES's future is occurring at a pivotal time of transition. The Futures Project was conceptualized under the leadership of the first Commissioner of Education Statistics, Emerson J. Elliott, and carried through under the stewardship of Jeanne E. Griffith as Acting Commissioner. I plan to use this publication in the upcoming years as a source of ideas for planning and thinking and as a foundation for long-term change in the organization.

Pascal D. Forgione, Jr.
Commissioner

Acknowledgments

From Data to Information: New Directions for the National Center for Education Statistics began more than two years ago as a proposal from Jeanne Griffith. With the strong support and encouragement of Emerson Elliott, then Commissioner of NCES, Jeanne and staff in the Data Development Division guided and nurtured the project to its successful completion in the spring of 1996. John Ralph played a key role in shaping the focus and direction of the project, with important early contributions from Ron Hall, Ed Mooney, Paul Planchon, and Joe Conaty. Edith McArthur and Dawn Nelson also helped shepherd the project along.

Eighteen authors wrote papers that provided the project's substantive foundation and then presented their work to a national conference held in November 1995. Discussants from both inside and outside NCES also participated in this conference and contributed their expertise. *From Data to Information* assembles the thought, creativity, and analysis of these authors and discussants and summarizes some important considerations for policy, research, and practice. Without these individuals, the project would not have been possible, and they deserve special appreciation for the care and enthusiasm they brought to the undertaking. A complete list of authors and discussants is included in each of the project's publications.

At MPR Associates, Phil Kaufman helped mold the policy and methodological directions of the papers and conference discussions. Steve Klein and Patricia Holmes conducted a survey of leading educators and researchers in order to gather their ideas about the essential topics to be addressed by the project. Barbara Kridl was responsible for overseeing the production of the final product and editing it, Andrea Livingston for designing the volume and editing the draft, and Karyn Madden for editing the volume throughout its various iterations. Leslie Retallick and Denise Bradby created the cover, and Don Eike and Connie Yin produced the final product. Special appreciation also goes to Fena Neustaedter who provided ongoing administrative support, and to Laura Horn who offered many useful comments on the synthesis of the papers and conference proceedings.

To all of these individuals, thank you. Your work will figure prominently in discussions at NCES about both its current and future efforts to better inform the nation about education.

Gary Hoachlander
Project Director
June 1996

Contents

Page

Commissioner's Statement . . . iii

Acknowledgments . . . v

1 **From Data to Information: New Directions for the National Center for Education Statistics**
From Data to Information: New Directions for the National Center for Education Statistics
Gary Hoachlander . . . 1-1
Introductory Comments
Emerson Elliott . . . 1-28

2 **Tracking Education Reform: Implications for Collecting National Data Through 2010**
Tracking Education Reform: What Type of National Data Should Be Collected Through 2010?
John F. Jennings and Diane Stark . . . 2-1
Where Are We Going? Policy Implications for Data Collection Through 2010
Christopher T. Cross and Amy Rukea Stempel . . . 2-12
Discussant Comments . . . 2-19

3 **Curriculum, Pedagogy, and Professional Development**
Enhancing Opportunity to Learn Measures in NCES Data
Dominic J. Brewer and Cathleen Stasz . . . 3-1
Teacher Education, Training, and Staff Development: Implications for National Surveys
David R. Mandel . . . 3-29
Discussant Comments . . . 3-43

4 **Trends in Statistical and Analytic Methodology: Implications for National Surveys**
"So What?" The Implications of New Analytic Methods for Designing NCES Surveys
Robert F. Boruch and George Terhanian . . . 4-1
Discussant Comments . . . 4-116

Contents (continued)

Page

5 New Data Collection Methodologies, Part II: Experimental Design

Incorporating Experimental Designs Into New NCES Data Collection Methodologies
Charles E. Metcalf 5-1
Discussant Comments 5-19

6 Postsecondary Education

Tracking the Costs and Benefits of Postsecondary Education: Implications for National Surveys
Michael S. McPherson and Morton O. Schapiro 6-1
Special Issues in Postsecondary Education and Lifelong Learning
David W. Breneman and Frederick J. Galloway 6-13
Discussant Comments 6-29

7 New Data Collection Methodologies, Part I: Observational Strategies

Large-Scale Video Surveys for the Study of Classroom Processes
James W. Stigler 7-1
Discussant Comments 7-30

8 Education for Work: Curriculum, Performance, and Labor Market Outcomes

Education and Work: Curriculum, Performance, and Job-Related Outcomes
Peter Cappelli 8-1
Discussant Comments 8-35

9 Using Administrative Records and New Developments in Technology

Administrative Record Opportunities in Education Survey Research
Fritz Scheuren 9-1
New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education
Glynn D. Ligon 9-32
Discussant Comments 9-66

Appendix A: About the Contributors A-1

Appendix B: Future NCES Data Collection Conference Agenda B-1

9

1

From Data to Information: New Directions for the National Center for Education Statistics

From Data to Information: New Directions for the National Center for Education Statistics

---

Gary Hoachlander

INTRODUCTION

In 1995, approximately 65 million Americans participated in elementary, secondary, or postsecondary education in the United States (U.S. Department of Education 1995b). Young people spent from one-fourth to one-half of their waking hours in school and school-related activities, while Americans of all ages continued to pursue some form of active learning that added to their repertoire of knowledge and skills. To serve these students, the nation's schools, colleges, and universities directly employed some 11.2 million people, even more than in the health industry. In addition, these direct services supported a substantial number of additional jobs in companies serving education through the production of everything from buses and computers to textbooks and software. All told, the nation spent more than $500 billion[1] on formal education, or approximately 7.4 percent of Gross Domestic Product (U.S. Department of Commerce 1995). Probably no other single activity has occupied so prominent a place in family, community, and working life.

The primary purpose of the National Center for Education Statistics (NCES) is to describe this education enterprise and inform the nation about it. Congress charges NCES with collecting and reporting "statistics and information showing the condition and progress of education in the United States and other nations in order to promote and accelerate the improvement of American education."[2] In doing so, the Center conducts a range of ongoing national surveys examining early childhood education, elementary and secondary education, postsecondary education, adult literacy, and the nation's libraries. Further, in cooperation with many other countries, it supports international surveys that aid in comparing educational progress and processes across nations. NCES carries out numerous analyses of these data and annually prepares more than 100 reports targeted toward policymakers, educators, researchers, and the American people.

Doing this job well would be necessary no matter what the focus, but when the subject assumes the magnitude and importance of education, this responsibility takes on special significance. Therefore, NCES must ensure that it continues to describe education fully and accurately and that it performs this function efficiently and thoroughly—in other words, that it remains well informed about important education issues, key advances in methods, and new developments in the technology of collecting, managing, analyzing, and reporting large amounts of information.

To this end, NCES undertook an in-depth examination of how best to direct its responsibilities for collecting and reporting information on education over the next decade. Three principal questions guided this effort:

1) What are the major issues and trends in education that NCES should aim to address through the first decade of the next century?
2) What are the most important advances in methods for collecting and analyzing information that should guide how NCES surveys are designed and used?
3) What opportunities do technological advances in data management and communications present for improving data collection and analysis and for disseminating findings and information effectively?

To help answer these questions, NCES conducted four activities: 1) a survey of leading educators and researchers, asking them to answer one or more of the three questions listed above; 2) commissioned papers addressing key topics suggested by the survey results; 3) a conference where the authors of the commissioned papers presented their work, with subsequent discussion by NCES staff and external reviewers; and 4) a published volume of the commissioned papers and discussants' comments. This paper summarizes and synthesizes the results of this work and consists of four major sections. This first introductory section provides a synopsis of the major themes and conclusions emerging from the papers and the conference. The second section describes the current foundation of NCES, delineating its core functions, operating principles, and program of work. The third summarizes some new directions that NCES could pursue to provide information for policy, research, and practice in American education, and also addresses some important methodological and technological opportunities. The paper ends with a brief conclusion.

Two dominant themes emerged from this collaborative effort. First, NCES must place greater emphasis on transforming raw data into information useful to policymakers, educators, researchers, and the general public than it does today. Accomplishing this goal will require that the relationships between NCES and data providers and between NCES and data users change significantly. During the next 5 to 10 years, the distinctions among these three parties—NCES, data providers, and data users—will become increasingly blurred, and their communications will probably become much more interactive, continuous, and two-way, with all three parties actively and simultaneously engaged in survey design, data collection, analysis, interpretation, presentation, and dissemination. Although technology will help pave the way for this transition, considerable conceptual thinking will also be required to take full advantage of the technological opportunities.

Second, in order to be more responsive to the demands for information about education, NCES will need to broaden its conception of what constitutes "data" and strategies for their collection. Traditionally, NCES has concentrated on designing and conducting surveys asking "closed-ended" questions that lend themselves to rapid, well-defined quantification. Although such surveys are likely to remain the hallmark of NCES's data collection activities for some time, the agency will need to pay more attention to how to supplement these data with various forms of "prequantified" material and observations. Technological developments will permit inexpensive collection of increasing amounts of textual, visual, and auditory data as integrated supplements to surveys. This capacity should make it easier for researchers to ask questions and explore subjects that they did not foresee when designing the survey, thus enriching analytic power and reducing the expense of designing and conducting new surveys to examine unanticipated concerns.

In addition to these two major themes, this effort led to five important conclusions about future directions for NCES. First, NCES should strive to produce information that addresses more immediate and specific policy concerns. While the agency's role in monitoring and describing major long-term trends in education must not be compromised, this role will assume even greater importance if the agency can also contribute in a timely way to more focused policy debates. The widespread emphasis on education reform during the past 10 years has spawned a large number of different strategies for improving education. As a result, policymakers at all levels—national, state, and local—want to know more about what has and has not been accomplished.

Second, surveys yielding better information that bears directly on the practices of teaching and learning would significantly enhance the contribution of NCES to both policy debate and research. Since current surveys produce scant data on the specific content of curriculum, the nature and frequency of discrete classroom activities, the practices of teachers, or the kinds of tasks students perform in order to learn, the classroom remains largely a "black box" that defies clear understanding and precise strategies for improvement. Without a clearer understanding of what constitutes effective classroom practices, it will be difficult to do more than simply describe what kinds of education reforms have been implemented. Whether they have, in fact, improved teaching and increased learning will remain unknown.

Third, survey designers should consider more carefully strategies that will permit integrated analysis of the interrelationships among education inputs, processes, and outcomes. Although existing surveys do an excellent job of providing nationally representative descriptive data on many important aspects of education, they do not, however, lend themselves very well to reliable causal analyses that might increase knowledge about what works and why. The descriptive power of national data must be preserved, but there are promising new designs emerging, which, if selectively incorporated into national surveys, might generate more robust conclusions about the relative effectiveness of various educational practices.

Fourth, NCES should make better use of data already collected and maintained by others. Doing so will help NCES simultaneously accomplish three aims: 1) expand the amount and type of data it collects; 2) adopt a wider range of data collection and analytic methods; and 3) function within the tight resource constraints that are certain to affect almost all federal agencies. Previously, NCES has pursued such a strategy with some success—for example, through the Common Core of Data (CCD) for elementary and secondary education and the Integrated Postsecondary Education Data System (IPEDS); however, high standards for data quality, especially comparability and reliability, have frequently forced the agency to collect new data that were already available in a somewhat different form or from a different time period. Without doubt, NCES must maintain its data quality standards, but increasing cooperation with states and localities, combined with rapidly improving data management technology and communications, should create opportunities for the Center to do a better job of streamlining and coordinating data collection.

Fifth, NCES will need to place increasingly greater emphasis on dissemination. Data and information are only as valuable as the breadth, quality, and timeliness of the uses made of them. Electronic storage media (data tapes and compact disks) and printed publications will surely remain the cornerstone of the agency's strategy for distributing data, tabulations, and the results of analysis. However, NCES should pay more attention to clearinghouse and brokerage functions, as well as effective use of electronic networks.

These themes and conclusions do not represent radical departures from the major path NCES has been pursuing in recent years. Indeed, as the following section illustrates, they are well suited to building on the foundation of core functions, operating principles, and programs of work that support the current agency. Nevertheless, serious attention to these ideas will almost certainly produce important differences in what the agency now does and how it does it.

BUILDING THE FUTURE ON THE CURRENT FOUNDATION

In 1986, the Panel to Evaluate the National Center for Education Statistics, a group created under the auspices of the National Academy of Sciences (NAS), reported on the results of its 2-year assessment of the mission and effectiveness of NCES (Levine, ed. 1986). The Panel was created to address the widespread perception that the existing agency had not yet developed "the image and the reality of a competent and objective major statistical organization serving the wide need for statistics about education in the United States" (Levine, ed. 1986, p. 13). To address such problems as quality of data, timeliness, conceptual obsolescence, and insufficient funding and staff, the Panel made many important recommendations, including the following:

- Clearly establish and define the Center's role in ensuring the availability of data needed to describe the condition of education in the United States;
- Improve the compilation of education program, staff, and financial data from the states, including developing closer collaboration with the states to ensure that the Center's program of work meets both NCES and state requirements for usefulness, relevance, quality, and reliability;
- Strengthen the Center's methodological and technical capacity through more systematic use of outside expertise in the Advisory Council of Education Statistics, as well as ad hoc advisory groups;
- Develop, publish, disseminate, and implement standards to guide all phases of the Center's work, including establishing an office of statistical standards headed by a chief statistician;
- In collaboration with the states, assess and improve the quality, consistency, and reliability of data obtained from state and local agencies, from institutions of higher education, and from other sources; and
- Institute a publications policy that clearly distinguishes between different types of reports—for example, statistical summaries and digests, analytic reports, descriptive reports, and reports on methodology—and develop a schedule of fixed release dates for selected key education statistics.

Ten years later, with the direction provided by the Panel, strong leadership at NCES, and support from Congress and the larger education community, NCES is much stronger and has become a widely respected statistical agency. The agency has significantly strengthened its core functions; operates under well-defined guiding principles and high standards for data collection, analysis, and reporting; and has established a clear program of work for reporting on the major aspects of education in the United States and other nations. Whereas 10 years ago, the future of NCES depended on rectifying fundamental weaknesses, today the agency's future can build on a strong foundation.

The Core Functions of NCES

The National Academy of Science's Committee on National Statistics defines the principal purpose of a federal statistical agency as "the compilation and analysis of data and the dissemination of information for statistical purposes" (Martin and Straf, eds. 1992). NCES adheres to this primary purpose by organizing its work around three core functions:

1) Survey Design and Data Collection
2) Information Production—data analysis, translation, and interpretation
3) Dissemination

Since national surveys are the primary means for NCES to collect data on education, during the past 10 years, the Center has devoted much effort to improving survey design and data collection. For instance, following the recommendations of the 1986 NAS Evaluation Panel, it has developed and implemented various strategies to improve data quality, to detect and reduce error, and to expedite data collection. The Center has also significantly improved the sophistication and efficiency of its sampling methods, increased its use of computer-assisted telephone interviewing, and systematically assessed the quality of data generated in its national surveys.

Moreover, NCES has strengthened and substantially expanded its capacity to analyze data. Rigorous statistical standards now govern all aspects of its analytic function, from simple tabulations to the most sophisticated multivariate analyses.[3] The Center routinely applies procedures for quality control to all of its surveys, which include analyzing data quality, eliminating unacceptable error, and producing methodological and descriptive summary reports before releasing survey data for public use.

Finally, NCES has greatly expanded and improved its dissemination function. Toward this end, the agency has developed and implemented publication standards that now guide the production of NCES reports and the release of public use data files. A central new feature of the agency's dissemination function has been developing strict policies for protecting the privacy of participants in NCES surveys. The Center not only applies safeguards to the data released to the public but also requires that its analytic contractors follow strict requirements for limiting access to prerelease data files and for maintaining the confidentiality of survey respondents. Failure to adhere to these requirements carries stiff fines, as well as the possibility of imprisonment.

Operating Principles

In carrying out these core functions, NCES adheres to three operating principles:[4]

1) Produce information that is policy relevant, while maintaining strict impartiality, institutional independence, and neutrality with respect to programmatic effectiveness;
2) Maintain credibility with users of its data, analysis, and publications; and
3) Maintain trust among those who provide data, including individuals, institutions, and public and private agencies.

NCES's program of work must be guided by the issues and requirements of public policy and federal programs, while scrupulously avoiding specific policy recommendations or identification with particular policy agendas or ideological perspectives. This principle is perhaps easiest to achieve when the Center performs its responsibilities for providing data to others for analysis or when it produces tabulations and descriptive reports. When NCES engages in analysis or interpretation, however, it must exercise greater care to remain policy neutral while still contributing relevant information to policy debates.

Attention to this principle has important implications for charting future directions for NCES. The suggestions that the Center address more immediate, specific policy concerns and that it develop survey designs more strongly suited for evaluation of what works could lead it beyond the boundaries of policy relevance into policy statements and evaluation. This, in turn, could jeopardize its position of impartiality. Deriving greater policy benefit from data and information produced by NCES, therefore, must proceed with great care.

Attention to this first principle also contributes to realizing the second, credibility with users of NCES data and information. However, credibility depends on more than policy relevance and impartiality. It also derives from confidence in the rigor of survey design, the quality of the data, the strength of analysis, and the accessibility and usability of its products, publications, and services. Here again, as NCES considers making greater use of data collected and maintained by others, it will need to guard against undermining its credibility with users who now depend on the Center's increasing emphasis on methodological rigor and data quality.

Finally, the success of NCES as an information agency rests on the trust it engenders among those who supply it with data. Protecting the privacy of survey participants is a key aspect of maintaining this trust, and integrating new types of data into national surveys will pose challenges for assurances of confidentiality. Use of video and audio data—for example, taping teachers in the classroom—will require close scrutiny of this issue. Confidentiality, however, is not the only condition for securing trust among data providers. Suppliers of data also need to be confident that the information being requested is truly needed, that it will be tabulated and analyzed accurately, and that providers will be given opportunities to correct errors or clarify ambiguities. Pressures for greater timeliness or more direct electronic access to decentralized, raw data files may undermine the confidence of data providers in the absence of explicit attention to new strategies and safeguards.

Program of Work

NCES has organized its current program of work around seven major topics:[5]

1) Elementary and Secondary Education
2) Postsecondary Education
3) Educational Assessment
4) National Longitudinal Studies
5) International Comparative Studies
6) Vocational Education
7) Libraries

Information on each of these topics is produced from a variety of surveys and studies, several of which supply data to more than one topical area. Some of the surveys, such as the CCD (on elementary and secondary schools and school districts), are designed as a census of the universe of respondents, which then serves as a sampling frame for more in-depth cross-sectional or longitudinal surveys on smaller samples of the population. The Schools and Staffing Survey (SASS), for example, collects detailed information on teachers and administrators in a sample of schools drawn from the CCD.[6] In other instances, a large comprehensive survey provides the basis for a more intensive study of a subset of respondents. In this vein, the National Postsecondary Student Aid Study (NPSAS)—a nationwide survey of students enrolled in postsecondary institutions—provides the basis for more targeted longitudinal studies of students who are starting postsecondary education, the Beginning Postsecondary Students (BPS) Longitudinal Study, and of students who have completed a baccalaureate degree or higher, the Baccalaureate and Beyond (B&B) Longitudinal Study.

Central to the NCES program of work are various surveys and studies designed to assess the knowledge, skills, and performance of American students. For instance, the National Assessment of Educational Progress (NAEP), which conducts assessments of reading, mathematics, writing, science, history, and geography for samples of students enrolled in elementary and secondary education,[7] is probably the best known of these efforts. In addition to NAEP, NCES also provides other data on student performance through transcript studies (at both the secondary and postsecondary levels); through the National Adult Literacy Survey, which examines adults' ability to use prose, documents, and mathematics in a variety of commonplace daily activities; and through international assessments that provide comparative information about student performance in the United States relative to that of other countries.

Finally, the Center conducts several long-term longitudinal studies designed to track students' paths through school and into subsequent stages of working and family life. These have included such studies as the 1980 High School and Beyond (HS&B Study), the National Education Longitudinal Study of 1988 (NELS:88), and the Early Childhood Longitudinal Study (ECLS), which is still in the planning and testing stage and is expected to begin with a kindergarten class in 1999.

These surveys now contribute to approximately 100 publications that NCES produces each year, including descriptive reports, analysis reports, methodological reports, issue briefs, and a variety of other documents. Three of these documents—the *Digest of Education Statistics, Projections of Education Statistics,* and *The Condition of Education*—annually provide a broad national overview of education at all levels in the United States.

In summary, during the past 10 years, NCES has been engaged in a process of steady development and improvement. In 1996, NCES is a viable and credible statistical agency, applying high standards to the provision of information on the condition of education in the United States and the nation's progress toward improving mastery of knowledge and skills among all its citizens. With this strong foundation, the agency is now well positioned to pursue some new directions that will enhance its ability to produce important information for policy, research, and practice in American education.

NEW DIRECTIONS IN INFORMATION FOR POLICY, RESEARCH, AND PRACTICE

In considering how NCES can best chart a course over the next decade that will capitalize on the foundation of work already in place, it is useful to consider its contribution to three domains of education: policy, research, and practice. These domains are by no means mutually exclusive; in fact, they overlap and interact in important ways. There are, however, information needs that are either unique or more dominant in each, and it is therefore instructive to consider the following questions individually:

- How can NCES best contribute information to discussions of education policy at the national, state, and local levels?
- How can NCES contribute information that will support significant research on education effectiveness and improvement?
- How can NCES contribute information that supports practice—i.e., the "front-line" activities that develop knowledge and skill in the nation's students?

A fourth question constantly weaves through these first three: what advances in methodology and technology can assist NCES in providing useful information to each of these domains? This section addresses each of these four questions in turn.

Information for Policy

The agenda of NCES is, in the first instance, greatly influenced by public policy issues and the requirements of federal, state, and local programs affecting education. Information contributing to policy debates can assume at least three major forms:

1) *System indicators* that describe the functioning of the education enterprise, broadly and over the long term;

2) *Implementation indicators* that describe the breadth and depth of the execution of policies and practices; and

3) *Effectiveness indicators* that describe the results achieved by students and educational institutions and programs.

Although system indicators have been a long-standing focus of NCES, there are potentially important new developments for the agency to consider. NCES surveys have included information on the implementation of some generic policies and practices, but specific federal and state policy initiatives have not been examined. Surveys have also included measures of student outcomes—the NAEP is the best known example; however, these measures typically cannot be directly linked to particular policies or educational practices to permit rigorous assessments of effectiveness. What are some possible new directions for NCES to consider with respect to each of these three types of indicators?

System Indicators

Data that portray the major aspects of the American education enterprise, both cross-sectionally and over time, form the core of the mission and functions of NCES. Reporting basic descriptive information on students, faculty and other staff, institutions and governing districts, and education finances must continue to be the primary focus of NCES and should not be compromised by new initiatives. The authors contributing to this examination of new directions for NCES are unanimous on this point: the primary purpose of NCES is and should remain representatively describing and documenting the condition of education in America and other nations.

This basic description of the education system can, of course, be improved, and several of the papers included in this volume offered suggestions.[8] Among the kinds of system information the authors would like to see developed are the following:

- Detail on curriculum content, including rigor and substance;
- Detail on the nature and frequency of particular teaching practices, especially those which research indicates are effective;
- Attention to the nature and frequency of student behavior that reflects engagement in learning;
- Resource allocation at the institutional and classroom level;
- Measures of teacher quality and the ways in which teachers apply their knowledge and skills in the classroom;
- More contextual information on postsecondary institutions, especially their objectives in awarding student financial aid and improved coverage of proprietary institutions;
- More attention to the interaction between education and work; and
- More attention to governance issues, particularly new organizational and oversight arrangements.

In many respects, these recommendations represent requests for "finer grain" in the descriptive data presently collected by NCES. In some instances, this aim can be accomplished by asking for more detailed information; in other instances, collecting and reporting existing data at lower levels of aggregation (the classroom, for example, rather than the school or school district) will be necessary.

Implementation Indicators

The widespread attention on education reform during the past 10 to 12 years has spawned a number of new policy initiatives at the national, state, and local levels. Congress periodically revises such mainstay education legislation as the Elementary and Secondary Education Act, the Higher Education Act, or the Individuals with Disabilities Education Act. Additionally, it undertakes new education policy initiatives such as GOALS 2000 or the School-to-Work Opportunities Act. States have also initiated many new policies to strengthen elementary, secondary, and postsecondary education. These have included changes in the requirements for high school graduation, new teacher certification regulations, modifications to postsecondary admissions standards, and new policies on college tuition and student financial aid.

Traditionally, NCES has not monitored the implementation of specific federal or state legislation. At the national level, Congress has typically provided for independent assessments or evaluations of education legislation, such as the National Assessments of Vocational Education, which have been conducted approximately every 5 years. While these national assessments make extensive use of NCES data, they also conduct independent surveys that focus more particularly on key features of the legislation being examined.

Several of the authors involved in this project have urged NCES to monitor some of the key policies and practices that have emerged from federal, state, and local legislation during the 1980s and 1990s.[9] It should be emphasized that they are not recommending that NCES assume responsibility for evaluating particular legislation, because they believe this function should continue to rest elsewhere. Rather, they are urging NCES to examine policies and practices that became more generic as they have been adopted and implemented through various federal, state, and local initiatives and are, therefore, no longer associated with any single piece of legislation. Some specific examples include the following:

- Curriculum content standards and measures of student or institutional performance;
- Length of the school day or year;
- Requirements that students complete particular courses (for example, in math, science, or foreign language) or accumulate a minimum number of credits for graduation;
- Participation in a variety of "work-based" learning opportunities, including apprenticeship, cooperative education, tech-prep programs, or school-based enterprise;
- Operation of charter schools;
- Prevalence and nature of home schooling;
- Availability and use of school choice;

- Participation in reform networks, such as the Coalition of Essential Schools, Accelerated Schools, or *High Schools That Work*;
- Changes in affirmative action policies;
- Changes in postsecondary admission requirements;
- Prevalence of state takeovers of local school districts or other forms of state intervention in financially troubled localities; and
- Changes in state policies affecting postsecondary tuition or student financial aid.

More attention to such issues by NCES would help ensure that its data are policy relevant, while still leaving responsibility for policy evaluation to independent studies and other agencies in the Department of Education.

Effectiveness Indicators

In addition to information on how to implement policies and practices, policymakers would also like better information on their effectiveness. Even though it is useful to know how widespread the adoption of a particular strategy for improving education has been, it is even more useful to know how well it has worked, and why or why not. This, of course, is a primary aim of most policy evaluation, as well as many research projects.

Much of the credibility of NCES rests on its clear separation from policy evaluation and research on education impacts and outcomes. Although NCES contributes essential data and information to these efforts, it remains well removed from the conduct of any of these activities. This separation of functions contributes to the neutrality and objectivity that NCES must maintain as the nation's primary statistical agency for education. The impartial character of NCES must be preserved. Consequently, any initiative to make NCES surveys more conducive to assessments of policy effectiveness must proceed with great care.

Why consider such a course at all? First, there is potentially a substantial payoff from better integrating the nationally representative features of NCES surveys with the more rigorous but also more narrowly circumscribed designs of policy impact studies. In the current environment, policy analysts often face a frustrating choice: asking the right question with weak methodology and data that were not collected specifically for that purpose, or asking a much less important question with sound methodology and specially tailored information.[10] Clearly, answering important questions with sound methods and precise information is more likely to improve education and the policies that support it. Combining the representative power of national surveys with the methodological rigor of experimental design would help realize this objective.

Second, there may be significant cost savings from integrating some impact evaluations with national surveys. Both kinds of efforts are quite costly. It is not unusual for a national survey to cost in excess of $10 million, and the more rigorous policy evaluations adopting experimental design frequently cost as much or more. Both efforts often collect similar kinds of data at approximately the same points in time, sometimes even from the same respondents.

Eliminating this duplication would not only reduce costs but also alleviate some of the burden on respondents participating in national surveys and evaluations.

Cost savings aside, the primary benefit of integrating methodologically rigorous effectiveness assessments with nationally representative surveys lies in increasing the usefulness of these two activities beyond the results obtained when they are conducted independently. National survey data would more directly and authoritatively address questions about policy effects; impact studies would be conducted in a nationally representative context that would increase the likelihood that study results could be generalized.

To achieve this result, NCES should carefully consider piloting the inclusion of an experimental study in one of its national surveys. Any of the longitudinal surveys now under way are potential candidates, including SASS, ECLS, or the longitudinal spin-offs of NPSAS.

What should be the focus of experimental studies imbedded or linked to national surveys? Clearly, the choice must be considered carefully, with ample input from interested policymakers, researchers, and educators. Given the mission of NCES, focusing on a particular type of educational practice would probably be more appropriate than on an assessment of a specific legislative program. One possibility, for example, would be to conduct a careful study of the consequences of homogeneous versus heterogeneous grouping of students by academic ability.[11]

Information for Research

Researchers are heavy users of information produced by NCES, and while the boundaries between policy and research are fuzzy, the interests of the research community deserve some separate attention. In the papers produced for this project, three themes emerged as priorities for focusing NCES's contribution to research over the next decade:

1) Teaching and Learning—illuminating more clearly what actually happens in the classroom;
2) Education Production—clarifying the processes of transforming education resources into student, program, and institutional outcomes; and
3) Education Outside the Classroom—depicting what and how learning occurs beyond the walls of the traditional classroom in homes, workplaces, and the community at large.

Though not exhaustive, this list provides some important directions for NCES to consider. In the next section, each topic will be briefly discussed.

Teaching and Learning

Much of the business of education occurs in the nation's classrooms—elementary, secondary, and postsecondary—yet national surveys presently tell us relatively little about what actually takes place at the classroom level. Currently, good information is available about the different types of courses taught (at both the secondary and postsecondary levels), but there is little or no nationally representative detail on the content of the curriculum or how it varies among classrooms, institutions, or states. Similarly, not much data are available on teaching practices, either the range of strategies adopted by faculty or the frequency of their use. Finally, most surveys do not offer much description of what students do to facilitate or impede learning in the classroom.[12]

While richer information on these three aspects—curriculum, pedagogy, and student behavior—would be useful in and of itself, the greatest benefit to research is likely to be achieved when information on all three is simultaneously available at the individual classroom level. That is, ideally researchers would want to examine how these three aspects of classroom activity interact and to understand how they relate to various types of education outcomes. In this way, more can be learned about what works and why in the daily business of education.

To realize this objective, one implication for future NCES surveys is clear: survey designs need to pay more attention to using the classroom as a unit of analysis. Additionally, the designs should strive to produce an *integrated* package of information on curriculum content, teaching practices, student behaviors, and student learning outcomes. It is not sufficient, for example, to simply expand transcript studies to include more information on course content; rather, expanded information on course content must be linked to other data on teaching practices, student behavior, and student achievement.

In addition to the question of what kinds of information on classroom activity can best advance future research, there is also the issue of how best to collect it. Traditionally, to obtain data on classroom activities, NCES has asked respondents questions through paper questionnaires or telephone interviews. Thus, to the extent that current surveys yield information on teaching practices or student behavior, they rely mainly on self-reports.

An alternative to collecting information through respondent self-report is direct observation by trained researchers. Until recently, direct observation has been a very expensive alternative, indeed prohibitively so for large-scale surveys involving thousands of respondents. However, recent technological and methodological advances are making direct observation, as well as the collection of source materials, more feasible.[13] Video is one of the most promising strategies for linking direct observation to more traditional survey techniques, and NCES is using this technique for the first time in designing the Third International Math and Science Survey (TIMSS).[14]

Video, of course, is not an especially new technology. What is new, however, is its rapidly growing capacity to store large amounts of video information inexpensively in digital form that enables fast retrieval and analysis. Additionally, researchers are making steady progress in developing analytic techniques that simplify and accelerate transforming video information into coded data suited for analysis using quantitative methods. Video, therefore, can add significantly to the richness and analytic potential of a survey, since it reduces the need to anticipate all of the questions the survey must ask of respondents. As researchers observe video records, they can formulate completely new variables that may not have been considered in the design phase of the survey. In the past, such new formulations usually required asking respondents follow-up questions or designing a new survey, if such avenues were pursued at all.

Closely related to this kind of use of video technology is the increasing capacity to collect, store, and analyze large amounts of textual information. For example, if researchers need better information on the content of textbooks or other printed materials used in classrooms, it is now possible to optically scan samples of these classroom documents for subsequent coding and analysis. As with video images, electronically storing and retrieving large amounts of textual information is relatively inexpensive.

These advances in storing and analyzing large amounts of what is essentially "prequantified" data promise to integrate survey research with case study methods, and represent research strategies that until now have been pursued independently of one another, each with its own strengths and weaknesses. This integration has the potential to link the representative statistical power of survey design with the richness and variety of case study information, simultaneously obtaining the best of both worlds.

Expanding surveys to include systematic collection of prequantified visual, textual, and even auditory information could produce significant new contributions to research on teaching and learning. Consequently, NCES should carefully consider how best to capitalize on its initial experience with this strategy in TIMSS, with special attention to adapting the approach to other surveys such as SASS, ECLS, or the longitudinal spin-offs of NPSAS. Additionally, the use of video in national surveys might prove especially beneficial if it were initially combined with efforts to imbed experiments in national surveys. The combination of these two methods targeted on analyzing the effectiveness of particular teaching interventions, for example, could yield very useful and robust results.

Education Production

Better understanding the interactions among curriculum, pedagogy, and student behavior in the classroom is an important piece of a larger set of research questions—how dollars are allocated (to localities, institutions, and classrooms), transformed into various resources, organized into programmatic and teaching strategies, and used to produce increases in students' knowledge and skills. In short, NCES data could play a much more significant role in expanding knowledge about how to better use education resources to improve student performance, thereby improving the overall process of education production.[15]

Achieving this goal will require some changes in the way NCES currently collects data on the financing of elementary, secondary, and postsecondary education. At present, there are two main surveys collecting financial data, CCD at the elementary and secondary level and IPEDS at the postsecondary level.[16] With respect to financial information, both of these surveys focus primarily on providing detail on revenues and expenditures, for local school districts in the case of CCD and for individual institutions in the case of IPEDS. Both surveys are designed to collect financial data primarily from an accounting perspective and are not now well suited for cost-benefit analysis of educational programs or cost-effectiveness analysis of particular teaching strategies. Neither provides information on the allocation of resources at the classroom level.

Providing data that better inform understanding of the production process of education would be aided by NCES expanding its present focus on finance to embrace a broader concentration on the economics of education. This larger conception would aim to integrate data on finance with other data on education processes and practices, as well as student outcomes. Additionally, new kinds of economic data would be required. Rather than needing more detail on expenditures for such general functions as administration, instruction, maintenance, or capital outlay, researchers would want to obtain data on the costs of specific types of staff, different kinds of school improvement strategies, alternative teaching strategies, and so on. They would also want to know more about the costs of different kinds of course content, equipment, instructional products, and assessment. In short, rendering the process of education production more intelligible depends on moving beyond traditional concerns about the distribution and expenditure of dollars to a more careful examination of how to transform dollars into effective teaching and learning in the classroom.

Three strategies for improving NCES data on education finances would help accomplish this goal. First, what constitutes useful financial data needs to be reconsidered, with special attention to better information on unit costs and transforming dollars into education processes and practices. Second, data will be needed at the classroom level; information on districts or institutions is not likely to contribute much to this kind of research. Third, it must be possible to link these financial data to other data on teacher characteristics, classroom practices, student demographics and behavior, and learning outcomes. Without this kind of integrated data about how education occurs, understanding more precisely how to efficiently allocate resources for education will continue to elude researchers and policymakers.

Education Outside the Classroom

Although elementary, secondary, and postsecondary classrooms are the centers of formal education in America, it is widely understood that much learning also takes place outside the classroom in the home, the workplace, and the community at large. However, we know relatively little about what or how learning occurs in these settings, nor do we know much about how learning in these places interacts with learning in the classroom. Moreover, given that most Americans spend only 12 to 16 years in formal schooling but another 50 years or so learning in these informal environs, a thorough description of the condition of education in America would require closer attention to the learning that transpires beyond classroom walls.

NCES surveys have already paid some attention to nonschool settings. At the present time, probably the largest of such efforts is the Early Childhood Longitudinal Study (ECLS), which will begin by focusing on the preschool lives of a cohort of children who will be followed over their early years of development. Additionally, other longitudinal studies, such as HS&B and B&B, have collected data on respondents' experiences in the workplace. Information on labor market participation, however, has been limited primarily to data on types of labor market outcomes—for example, earnings, duration of employment, and types of occupation—rather than systematically examining how learning occurs in the workplace or the degree of congruence between learning goals in schools and education requirements on the job.

Comprehensively surveying learning that occurs outside the classroom is a tall order, and NCES should approach this task incrementally. One place to focus an expanded examination of informal learning is on the workplace and the strategies adults use to maintain and upgrade the knowledge and skills needed to remain productive, actively engaged workers.[17] Such a focus is more important than ever, given the changes that are occurring in today's work world. These changes include not only rapidly developing new technologies but also major shifts in the attachments and relationships between employers and employees. As the likelihood of lasting employment with a single employer becomes increasingly tenuous in the modern economy, individuals must assume ever greater responsibility for nurturing their own careers and continuing employability. How working adults will meet this responsibility in the future poses important new challenges for the nation's systems of education.

Increased attention to learning through and for work could begin with the following steps. First, it is important to learn more about the knowledge and skills needed for long-term success in the labor market. Are these requirements consistent with the academic and vocational goals of formal education, and how well does the formal education system produce the desired prerequisites? Second, NCES could pay closer attention to how learning occurs in the workplace; whether the process differs in important ways from learning in the classroom, and whether the two complement or reinforce one another. Third, NCES could enrich work-related data in its current longitudinal surveys, concentrating especially on better information about what people do on the job, what contributes to their successes and failures, and how they use or do not use school-based learning to perform and advance.

As part of its own mission, NCES could independently address all three of these issues. Alternatively, the agency may want to explore opportunities for collaborating with the Department of Labor and its surveys of employers and employees. Data collected by the Labor Department tend to provide greater detail on labor market participation, while being relatively deficient on education variables. Better coordination or integration of the two Departments' survey efforts could yield some important benefits.

Information for Practice

Most teachers and administrators are accustomed to viewing data as something to be reported to others. For example, they report daily attendance to central offices to document federal, state, and local funding systems. They submit grades for report cards to students and parents and for recording on student transcripts, which in turn are reported to postsecondary admissions offices. They administer standardized achievement tests for state assessments and college admissions. Rarely, however, do teachers and administrators use data directly themselves to improve their own programs and practices. One consequence of this outlook on data is that most practitioners do not make much use of the information provided by NCES. An important challenge for NCES, therefore, is significantly increasing the value and utility of its data for local teachers and administrators.

At least three strategies for providing better information for practice offer some important opportunities for NCES:

1) *Benchmarking*—helping local practitioners make comparisons against established norms;
2) *Networking*—linking practitioners with other practitioners and helping them discover more quickly who is doing what and where;

3) *Practitioner-Based Research and Self-Reflection*—engaging practitioners in systematic inquiry through NCES surveys and related research.

Benchmarking

"How well am I performing?" This is an appropriate question for any professional concerned with improving practice and increasing students' mastery of knowledge and skills. For most educators, however, it is a difficult question to answer in any way other than in an impressionistic or anecdotal fashion. Until recently, education has not had enough success with helping schools, programs, and faculty to monitor their accomplishments or to use the results to improve what they do.

Fortunately, as more and more states and localities develop new strategies for tracking performance and promoting school improvement, this situation is changing. "School report cards" are now produced annually in many states. Other states have developed systems of performance measures and standards, along with procedures for school improvement plans in districts that perform below state norms. Moreover, "keeping score" and using the results to assess the relative effectiveness of different kinds of school improvement strategies are core operating principles of several large consortiums, such as *High Schools That Work* under the auspices of the Southern Regional Education Board.

NCES could make an important contribution to the continued development of these practices by improving the utility of its survey results as benchmarks for states and localities interested in knowing how their performance measures up in relation to others. A local school or school district, for example, could find out how well its record on student attendance or high school completion compares with a national or state norm. It could then further refine the comparison by examining such measures in a subset of districts or schools that are similar with respect to size or student demographics. In addition to making comparisons at a particular point in time, a local school or school district might also monitor its relative performance over time. For example, is its success in reducing dropout rates proceeding at a faster or slower pace than in comparable districts or schools?

There is nothing to prevent localities, or even individual teachers, from using current NCES data to establish these kinds of benchmarks. However, they need to work rather hard to do it. Finding the right data is not always easy, nor is determining whether the NCES estimate is comparable with a local statistic. Tailoring an NCES estimate to yield a comparison of "likes with likes" requires a knowledge of NCES data sets, as well as analysis techniques, that most practitioners do not have. Thus, there are significant barriers to transforming NCES data into useful benchmarks at the local level.

However, there are at least three steps that NCES could take to make benchmarking easier for states and localities. First, in collaboration with potential state and local user-practitioners, NCES could systematically review its current dissemination activities with specific attention to how some aspects of the dissemination process could be modified to facilitate benchmarking. For example, there may be consensus on a relatively small set of indicators that NCES could publish annually in a succinct, accessible form with widespread local distribution. Such a publication might be similar to *The Pocket Condition of Education,* which NCES now produces annually, but it could be designed with local benchmarking specifically in mind.

Second, NCES could explicitly consider local benchmarking when designing selected surveys, including customized reporting of results to survey participants. At present, survey participants receive little or no direct benefit from taking part in NCES surveys, and the burden of doing so is often not trivial. Providing participants with a summary of where they stand on selected variables relative to others in the survey could be a useful service.[18] Such a summary could take the form of a traditional printed report. Alternatively, NCES might want to explore new electronic strategies that could actually distribute some limited analytic capacity along with the data (see discussion below on technological innovations).

Third, as NCES increases its capacity grows to provide information "on-line," it should consider strategies for developing and distributing analytic packages that enable state and local benchmarking. In other words, instead of simply making data available, NCES would also provide a menu of data analysis programs or routines that would enable practitioners to generate their own statistics quickly and easily, without requiring a sophisticated knowledge of the underlying methodology. Such a strategy would build on NCES's current practice of providing users with "table generators," increasing both the kinds of analysis that users could perform and the ease of using the analytic software.

Networking

Local teachers and administrators often want to know who else has experience with a particular school improvement strategy, type of curriculum, or teaching practice. Yet, systematically locating and communicating with other knowledgeable practitioners can be quite difficult; often it is not easy to find out who these individuals are or how to contact them. If NCES were to assume a greater role in monitoring the implementation of more specific education policies and practices (see earlier section on Implementation Indicators), it could also facilitate networking among practitioners. In addition to providing practitioners with information about the frequency with which a particular reform is being implemented and where it is being attempted, NCES could also match up interested parties and help them share information about their experiences.

This kind of knowledge brokering would represent a new function for NCES, one that may not be completely in keeping with traditional perspectives on the appropriate role of a statistical agency. Nevertheless, as NCES develops its presence on the Internet and the World Wide Web, this kind of service would be an obvious extension of its capacity to transform data into information valued by practitioners. Moreover, when providers of data also have a direct use for similar information from others, they are much more likely to respond to NCES's requests in an accurate and timely fashion. Thus, NCES's ability to monitor implementation for policy purposes could well be enhanced by its also using the information to provide an important service to teachers and administrators.

Practitioner-Based Research and Self-Reflection

Until recently, surprisingly little has been known about the specific elements of high-quality teaching (and by extension, high-quality teachers). This lack of knowledge has contributed to much misinformation and misunderstanding about what it takes to be a good teacher, as well as confusion in the public policy arena over the role of professional development in education reform. The status of national data on teachers reflects this state of affairs, with facts limited mainly to demographic characteristics and scant information available on the quality of practice or practitioners.

A very promising development, therefore, is the recent effort on the part of the teaching profession to begin a systematic, sustained examination of what constitutes good teaching—specifically what teachers should know and be able to do to help students master high levels of proficiency. Exemplified by the work of the National Board for Professional Teaching Standards (NBPTS), as well as other organizations and state-level initiatives, this effort is forging some consensus about appropriate standards for defining advanced high-quality teaching. This work has led to the establishment of a voluntary system of certification for early childhood, elementary, middle, and secondary school teachers, including differentiation among a range of academic disciplines (for example, math, science, history and the social sciences, English, and vocational education).

These developments create an important opportunity for NCES not only to improve the data it gathers on the nation's teachers but also to contribute more directly to strengthening teaching.[19] This opportunity can be realized in two ways. First, as the work on teaching standards and certification continues to evolve, NCES should be able to define a larger array of indicators of teaching quality to include in national surveys. At a minimum, these indicators should focus on measuring teachers' command of the knowledge bases and teaching methods that are being identified as reflective of high-quality practice. Gathering such data could rely on traditional methods of written assessment or self-report. Alternatively, if NCES opts for further developing video observation techniques, these methods could significantly enrich information on the condition of teaching nationwide.[20] Furthermore, as more teachers choose to pursue national certification and as more certificates are awarded, national counts of teachers participating in and successfully completing the process will assume greater value as indicators of teacher quality.

Second, as NCES pursues this first strategy for improving data on the quality of practice, it could actively engage practitioners in this process and create opportunities for more interactive research and development. For example, if written examinations (in the style of NAEP) or video observation become part of NCES's strategy for monitoring and reporting to the nation on teacher quality, this process could be designed to simultaneously benefit individual teachers participating in the surveys. This might be accomplished in several ways. The results of written assessments could be returned to individual teachers. Groups of teachers could be assembled to review and constructively critique video segments. Further, if the data gathering process included collecting mini-portfolios submitted by teachers, these could be systematically evaluated, with examples of best practices culled from the data and disseminated to teachers and teacher education institutions. In short, the business of collecting national data could begin to play a more direct role in professional development and the strategic improvement of teaching and schools.

Further Considerations About Methodology and Technology

Increasing the contribution of NCES to policy, research, and practice depends in part on closer attention to a number of methodological and technological opportunities. Some of these, such as imbedding experimental designs in national surveys or collecting prequantified data through video and optical scanning, have already been discussed. There are some additional considerations, however, that deserve special mention, including: 1) developments in using administrative records, 2) promising techniques for obtaining hard-to-get information or producing more finely tuned estimates, and 3) effective use of the Internet and World Wide Web.

Administrative Records

Much of the information sought by national surveys already exists, at least in an approximate form, in records maintained for administrative purposes by schools, postsecondary institutions, district offices, state agencies, and other public and private offices. Transcripts, for example, provide detailed information on courses attempted and completed, grades, credits earned, and scores from standardized tests. Personnel records contain data on teaching assignments, salaries, demographics, qualifications, and experience. And budget and accounting offices maintain extensive records on revenues and expenditures. To the extent that surveys can access and use these administrative records, they often can obtain information that is more accurate than the responses provided by survey participants, often at significantly less cost.

Several NCES surveys already rely heavily on administrative records for information. Some good examples are CCD, IPEDS, and the NPSAS. There are, however, two types of problems that have limited the usefulness of administrative records. First, the contents of the records may not meet acceptable standards of accuracy, consistency, and comparability. Second, access to administrative records is often problematic, for a variety of reasons ranging from concerns about confidentiality to technical problems that may be as mundane as locating the right filing cabinet in the right office.

Technological advances in computing and electronic networking promise to reduce both of these problems considerably over the next decade, and NCES should be alert to opportunities to exploit new developments.[21] First, electronic administrative records maintained in easy-to-use relational databases will increasingly become accepted practice among the nation's elementary, secondary, and postsecondary education systems,[22] since they will be building administrative databases to satisfy their own needs and uses for data. Consequently, *collecting* data will less often be viewed as an externally imposed burden and cost. Whether *providing* these data to national surveys will be seen as burdensome, however, will depend critically on the ease with which data can be transmitted to those requesting information. To facilitate transmission, NCES will need to pay particular attention to assisting with the standardization of data elements and with the development of cheap scannable forms and other strategies for promoting electronic access and transfer.

Second, as local educators and administrators become more sophisticated users of data (rather than just providers), the business of designing surveys, collecting and analyzing data, and reporting results is likely to become much more interactive. The traditional model in which NCES assumes primary responsibility for all of these functions is likely to yield to much more decentralized, distributed models in which the respective roles of surveyor and respondent become less distinct and more intertwined. For example, respondents who are also users of data may play a much greater role in defining survey questions and data elements. They may also develop specialized analyses (including analytic routines) that are shared with other respondent/users. NCES, in turn, may assume more responsibility for coordinating and brokering surveys, analyses, and reporting, rather than unilaterally directing and conducting all of these activities.

One possible implication of these trends is a reversal in the respective roles of independent surveys and administrative records in providing national data for education. To date, administrative records have mainly been adjuncts to large-scale surveys; they have supplemented data collected through written questionnaires or telephone interviews. In the not-too-distant future, administrative records may become the basic building blocks of national data systems, with smaller targeted questionnaires designed as supplements.

Methods for Producing Better Statistical Estimates

To provide good information for policy, research, and practice, NCES relies on a wide variety of survey design, data collection, and analytic methods—some relatively simple and widely known, others extraordinarily complex or reflecting recent advances in specialized fields. For purposes of this synthesis, a thorough discussion of survey and analytic methods is neither possible nor appropriate. The papers produced for this project, however, raised and discussed a number of methodological issues and developments. Four of these deserve special mention for careful consideration by NCES in charting its future course.[23]

First, as mentioned previously, NCES should exploit opportunities to combine well-designed, targeted controlled experiments within the national surveys that have been the hallmark of its data collection activities. These experiments must be compatible with the mission and conduct of the larger survey effort, and a particular experiment should not be undertaken if it risks jeopardizing the nationally representative and descriptive power of the survey in which it is imbedded. However, if these criteria can be satisfied, imbedded experiments are promising examples of constructing a "whole exceeding the sum of its parts." Such experiments could contribute significantly to knowledge about what works and why in the nation's classrooms.

Second, survey questions that elicit information on sensitive topics must always be considered with great care. National surveys about education are no exception, and it is important that NCES does not avoid issues simply because they are sensitive or controversial. Methodological developments can help reduce some of the concern surrounding this issue. For example, one promising strategy called "network-based estimating," in which respondents are asked about the behavior of unidentified acquaintances in their social network, has been developed by quantitative anthropologists. There is growing evidence that this procedure produces indirect but reliable information on sensitive topics, without depending on the respondent to report directly on his or her own personal experience. NCES should explore the feasibility of using this or similar techniques in future surveys.

Third, and related to the second issue, there are promising new developments in methods for generating indirect estimates of statistics at subnational levels or for intervening periods of time between surveys. Traditionally, producing estimates for smaller units of analysis—states, for example, or institutions within states—has depended primarily on increasing sample size. Similarly, obtaining estimates more frequently—say, every 5 years rather than every 10—typically requires administering the survey more frequently. Both of these strategies are usually quite expensive. An alternative method being developed uses auxiliary data (from ongoing administrative records, for example) along with the survey data to produce indirect estimates for smaller units of analysis or intervening periods of time. Successfully adapting these techniques to some NCES surveys could yield more finely grained estimates at a modest cost.

Fourth, there are long-standing calls for better linking and integrating the databases produced by NCES surveys. As noted earlier, a better understanding of how learning occurs in the classroom will require simultaneous access to data on curriculum content, teaching practices, student behavior, and student outcomes. It has been rare to find a single database with rich information on all of these attributes for a sufficiently large sample, however. If NCES were able to significantly improve the connections between its surveys, it is likely that opportunities for better, more focused research would be enhanced. However, in order to achieve this long sought-after objective, NCES must do substantial work, both conceptually and methodologically, to determine precisely what is meant by "linking" and "integrating."

Finally, NCES must continue to actively promote methodological developments and adaptations suited to its mission. Most researchers, whether engaged in a particular substantive pursuit or methodological advancement, are occupied primarily with their own interests and agendas; they are not paying much attention to the relevance of their work for NCES. Consequently, NCES needs to provide for the orderly acquisition and screening of methodological and technological applications to surveys and analysis. There are many strategies for doing this, including advisory groups, grants, conferences, commissioned papers, and so on. Whatever strategy is chosen, however, the basic objective must be an explicit and high-priority item on the agenda of NCES.

Internet and the World Wide Web

No discussion of future technological developments would be complete without some mention of the Internet and the World Wide Web.[24] However, the pace and variety with which these are evolving make any effort to forecast precisely their role in the work of NCES quite difficult, if not simply foolish. Perhaps, the most useful approach is to use the evolution of the information highway as a metaphor for changes in communication and interaction between NCES and the public it serves. In some respects, the Internet and the World Wide Web will facilitate and hasten these changes, but in others, they are simply reflective of larger forces at work in contemporary society.

Presently, NCES is in the second stage of a three-stage evolution in how many organizations typically interact with their clients. In Stage One, to accomplish its mission, NCES dominates the relationship between itself and those who either provide or use the data it gathers. Communication tends to be mainly one-way and follows well-established paths. NCES designs the surveys, administers questionnaires or interviews, collects and cleans data, conducts its own analysis, and produces and disseminates reports. Other analysts of NCES data pursue their research independently; they do not feed back results to NCES, at least in any systematic fashion. Stage One loosely represents the pre-Internet, pre-Web world. It is history.

In Stage Two, which coincides with the advent and initial development of electronic networks, relations between NCES and providers and users of data become more interactive (though still predominantly one-way), and in some instances the distinction between data *provider* and *user* begins to blur. In this stage, surveys begin to make more use of electronically stored administrative records, and, consequently, questions or data elements may be tailored to a particular respondent. Selected providers and users may be authorized limited on-line access data to update or correct information. Although NCES continues to generate substantial analyses on its own, it also begins to pay closer attention to the analytic objectives of users. In addition to distributing data files, it also disseminates analysis files designed to facilitate specific types of research—the relationship between education and labor market participation, for example. Additionally, NCES may provide analysts with software to accelerate their analyses or to ensure that those who conduct external analyses of NCES data adopt appropriate statistical techniques. Reports are made available in electronic form, and specialized electronic user groups or technical review panels begin to form on the network. Currently, NCES is already well immersed in Stage Two.

In Stage Three, which will emerge more clearly and strongly with greater access to electronic networks and with deeper understanding about how to use them effectively, relationships between NCES and its data providers and users will become truly two-way and continuous. Any data user, who could also be a data provider (a state office, for example), might send NCES a small software program that initiates a customized database search, adds the results to NCES's data library, and returns a tailored report to the original requester. Conversely, NCES may be constantly developing small software programs that go out over the network and retrieve data needed to respond to specific inquiries from Congress, researchers, educators, or the public at large.[25]

Surveys may assume the form of database development, with specifications designed interactively by users and providers coordinated by NCES. The scale of written questionnaires or telephone interviews will diminish considerably or be limited to highly focused inquiries. Much of this design process will occur on-line through electronic conferencing among NCES, data users, and data providers. Even though NCES will probably still produce many of its own reports, electronic versions of these documents will contain numerous electronic links to other data sets, technical references, and related reports. They may also contain interactive software that will permit users to perform "what if" analyses while perusing a report and to generate customized tables or graphics. Alternatively, users will generate their own electronic reports and analyses and create links to NCES documents residing in electronic networks. Precisely what is generated by NCES and what is generated by others may become less easy to distinguish.

Stage Three is not here yet, and it will probably look quite different from this admittedly inchoate prediction. However, this stage will probably arrive much sooner than expected. The more NCES can anticipate and help shape these developments, the more likely it will be able to use them effectively to report on the condition of education in the United States and other countries.

CONCLUSION

From data to information—transforming quantitative facts about education into knowledge useful to policymakers, researchers, practitioners, and the general public—this aim has always been central to the mission of the National Center for Education Statistics. In and of itself, this objective is not a new direction for the agency. However, what constitutes useful information and how it gets produced, distributed, and used are changing. To keep pace with these changes, indeed to stay out in front and help shape their development, NCES must chart some new directions.

Probably the most fundamental change that NCES will need to address is its emergence as a provider of information *services and systems,* rather than a primary collector and provider of data per se. In today's climate of growing demands for information, but limited resources to produce it, NCES will need to pay particular attention to assuming new roles as a facilitator, broker, translator, linkage, filter, and pathfinder in a complex web of providers and users of education data. To these new roles, the agency can bring a strong foundation of standards for high-quality data and analysis, as well as a firm understanding of the kinds of information that are most relevant to deliberating national policy for education.

As these new roles develop, NCES may find itself shedding or at least de-emphasizing old functions. Data collection that occurs independently of front-line administrative and teaching systems and their own information needs is likely to diminish significantly. This change, in combination with technological advances, may lead to data collection systems that are far more decentralized, interactive, and operating in "real time" than the systems that have traditionally supported national surveys. It is even possible that eventually NCES may find that it is no longer in the data collection business, as this function has traditionally been defined. Instead, it will be primarily a systems manager and analyst, a producer and broker of information for ongoing nationally oriented assessments, as well as thousands of state and local customized queries. Data collection and storage, however, may occur largely outside of the immediate domain of NCES.

"Reporting statistics and information showing the condition and progress of education in the United States and other nations in order to promote and accelerate the improvement of American education"—this charge is a lasting mission for the National Center for Education Statistics. Fulfilling it successfully will require careful attention to changing national priorities, a strong commitment to improving education research and practice, and an openness to recognizing and adopting important advances in methods and technology.

NOTES

1. U.S. Department of Education, National Center for Education Statistics, "Common Core of Data" and "Financial Statistics of Institutions of Higher Education," surveys and unpublished data, FY 94–95.

2. Section 402(b) of the National Education Statistics Act of 1994 (20 U.S.C. 9001).

3. Indeed, there are now standards for all of NCES major activities—survey planning and testing, statistical processing, data provision and analysis, evaluation and documentation, and contract management and operations. See U.S. Department of Education (1992).

4. These follow the principles for a federal statistical agency developed in Martin and Straf, eds. (1992).

5. For a full description of the current program of work at NCES, see U.S. Department of Education (1995a).

6. SASS also surveys private school teachers in a sample of schools drawn from the Private School Universe file maintained by NCES.

7. At present, legislation requires that NAEP assess reading and mathematics every 2 years; science and writing at least every 4 years; and history, geography, and other subjects determined by the National Assessment Governing Board at least every 6 years.

8. In particular, see the papers by Brewer and Stasz, Mandel, McPherson and Schapiro, Breneman and Galloway, and Cappelli.

9. See especially the papers by Jennings and Stark and by Cross and Stempel.

10. See the papers by Metcalf and by Boruch and Terhanian.

11. See the discussion in the paper by Boruch and Terhanian. These authors also suggest that NCES consider adopting a "satellite" policy that would permit including controlled experimental studies in national surveys in a fashion similar to the way NASA allows adjuncts to space missions for astrophysicists and others.

12. See the papers by Brewer and Stasz and by Mandel.

13. For a summary of trends in technological capacity to store, retrieve, and analyze data, see the paper by Ligon.

14. See the paper by Stigler.

15. See the paper by McPherson and Schapiro.

16. While SASS provides additional data on salaries of administrative and instructional personnel at the elementary and secondary level, NPSAS provides additional postsecondary information on tuition and costs.

17. See the paper by Cappelli.

18. There are important confidentiality considerations that must be addressed if this kind of service were to be provided. However, with explicit attention to benchmarking at the outset of a survey, problems surrounding confidentiality could be reduced.

19. See the paper by Mandel.

20. Video already plays an important role in the certification process used by NBPTS, and NCES could build on the experience of the Board, as well as that of other researchers developing this technology.

21. See the papers by Ligon and by Scheuren.

22. Electronic recordkeeping is still far from universal, especially at the elementary and secondary levels; paper files are still the norm in many places.

23. These are developed in more detail in the papers by Boruch and Terhanian, Metcalf, and Scheuren.

24. For more information about specific opportunities for NCES to use electronic networks, see the papers by Boruch and Terhanian, Ligon, and Scheuren.

25. The continuing development of "object technology," a technique for more rapidly constructing software programs out of many small object modules, should hasten the explosion of this sort of interactive dissemination and sharing of programs.

REFERENCES

Levine, D. B. (Ed.). 1986. *Creating a Center for Education Statistics: A Time for Action.* Washington, DC: National Academy Press.

Martin, M. E., and Straf M. S. (Eds.). 1992. *Principles and Practices for a Federal Statistical Agency,* p. 2. Washington, DC: National Academy Press.

U.S. Department of Commerce. 1995. *Statistical Abstract,* table 699, p. 451. Washington, DC: Bureau of the Census.

U.S. Department of Education. 1992. *NCES Statistical Standards.* Washington, DC: National Center for Education Statistics.

U.S. Department of Education. 1995a. *Programs and Plans: 1995 Edition.* Washington, DC: National Center for Education Statistics.

U.S. Department of Education. 1995b. *Digest of Education Statistics 1995,* table 3, p. 12. Washington, DC: National Center for Education Statistics.

Introductory Comments

Emerson Elliott

The authors of papers described in this volume have performed their services to the National Center for Education Statistics thoughtfully and with a serious purpose. As a collection, these examinations of possible future directions for education statistics are valuable—although not in laying out a plan of action for the Center, because the Center's leadership must do that. Rather, their value is in sketching a vision for a federal statistical information service and in showing that new technologies, statistical methodologies, and more powerful analytic procedures can achieve that vision even in a time when government is under pressure to achieve more with less.

Why is that so special? It was my frequent experience, during the years I was with NCES, to solicit advice about what the Center should be doing through commissioned papers or conferences. What data should it gather? What were the issues about which policymakers, educators, and the public needed information? And precisely what information from which sources (e.g., students, teachers, parents, schools, school boards) would answer those questions? Center staff have made aggressive efforts to keep abreast of technology in data gathering and analysis, and following criticisms from a National Academy of Sciences evaluation in 1986, the Center made numerous changes to assure that appropriate methodologies and rigorous quality controls were being employed in its statistical work. But those experiences, together, led me to conclude that the Center's now expanded budget and program activities had reached their limit in recent years. These new papers, though, while certainly assuming at least modest continuing growth in the Center's program, also indicate a potential for achieving much more statistical information for the investment.

My comments first address the sense of vision that flows through these pages and then turn to three implications of that vision—on technology and administrative records, research and statistics, and analysis.

VISION

The major theme expressed in the papers is that the Center should serve as an education information agency, not just a statistical office. This perspective is especially prominent in the papers prepared by Jack Jennings and Diane Stark and by Christopher Cross and Amy Stempel, although it is implicit in the others as well. These papers described an agency responsive to policymaker interests in education information about a variety of current topics, that would be gathered in different forms (some of them costly, such as longitudinal studies) and that would make the resulting information available in accessible forms. The Center would, among other things, be a source of information about the progress of reform even if that information is descriptive and not "statistical" in either the random survey sense or in the systematic extracting of data from traditional administrative records.

For example, Jennings and Stark emphasize in their paper that the Center should provide information on standards-based reform, on charter schools, on state course-taking requirements, on "reform networks" (such as those of Levin and Slavin), on choice programs, and on private, for-profit, companies that provide educational services.

Cross and Stemple, similarly, according to the Center should provide data on student mobility, school safety, moral and character education, technology, home schooling, charter schools, magnet schools, vouchers, and site-based management.

Thus, Jennings and Cross and their colleagues cast the Center not as a "data collection agency" but as a federal office with a broader education information function. It reminds me of the decline in the American railroads following World War II. Railroad executives viewed themselves as being in the railroad business, and lost out as transportation for individuals came to be defined by interstate highways, widely affordable cars, new airports, and jet engines. They never recovered from their own restricted vision of the railroad business.

Data collection is a tool that provides information. But that is too narrow a conception as questions turn to what the data say. The challenging question of what the data *say* can only be answered through an analysis function, through decisions about specifically which data will tell us something important, and through attention to relationships in data that have been examined and explained in research. Moreover, the very meaning of "data collection" is changing as electronic data systems come into widespread use, creating the possibility of moving information from administrative records instantaneously from place to place. The new potential of electronic capability is much more information at little cost, including data from "small areas" (schools, even classrooms, rather than districts or states or the nation) that are difficult and expensive to reach with traditional statistical methods.

At the same time, realizing this new vision would impose a change on the role of NCES as a statistical agency. The Center would "own" less data, since the information would be distributed around the nation at sponsor sites, frequently school districts, states, or institutions of higher education. The Center could capture the data for summarizing and analytic purposes, but others could as well—and surely would. As Glynn Ligon put it, the line between data retrieval and data dissemination would be blurred. The Center's role in building consensus for terms and definitions and common means of access, already exercised, would grow, probably with support and encouragement from the nation's educational and governmental institutions. They, too, will want access to data in some predictable way and will find the resulting statistics of greater use if they have the same meaning from place to place and can be connected from elementary and secondary education to higher education.

This vision is very attractive because it would make the Center relevant, I think essential, in the 21st century. It would, however, require developing connections and consensus about what is useful that involve more partners more consistently than has been the practice up to now.

ON RESEARCH AND STATISTICS

This group of papers describes roles for the National Center for Education Statistics that would require activities well beyond the bounds of traditional statistical agencies. Let us start with some brief definitions. I think of "statistics" as data that are designed to respond to those instances when information is to be representative of a population—when you want to know how something is distributed among individuals in different situations; when you want to know what happens over time; and when you want to know if a relationship derived from theories, experiments, and case studies will hold up when you "go to scale." "Research" also produces data, but is driven by a need to understand qualitative relationships, a need for in-depth information, or an intent to evaluate consequences of a specific intervention. There is no reason that policymakers should be required to make these distinctions—that should be done right here in the Office of Educational Research and Improvement. And it is certain in this collection of commissioned papers that the authors did not attempt to sort out what NCES should do and what other parts of the government should do.

Nonetheless, even a cursory review of the examples quoted above from Jack Jennings and Chris Cross—and more could be cited from other papers—will make it clear that just counting something (e.g., charter schools, number of students who move, states with choice programs) will be of limited value. A case in point is that, as Dominic Brewer and Cathy Stasz remind us, in tracking reform, change takes time and it is problematic just finding appropriate words for survey items that adequately describe what should be taking place. Even so, they note, surveys can only provide a snapshot of what is happening at one point in time.

Many of the specific suggestions for NCES data in these commissioned papers call for information about conditions in education that need measurement, such as instructional processes, curricular offerings and content, and the act of teaching. Where the intent of that measurement is to obtain information of a qualitative character—such as describing whether student content standards are equivalent, say, to those of NCTM in math or of NRC in science, or describing effects of instructional changes that can be associated with new student standards, and especially evaluating "educational treatments"—more analytical tools would be required and more micro-level data and case studies that are not feasible for national statistical programs would be needed. The Brewer and Stasz paper reminds us that theories should drive such data collections.

The need to ground data in compelling research theories and findings makes links with research crucial in any field of government statistics activity. I have recently chaired a review of the Joint Program on Survey Methods, an NSF-funded project housed at the University of Maryland, to train federal statistics staff. One of our panelists, an economist, insisted that the training of government statistics staff could be adequate only when statistical training was informed by the linking of statistical planning, design, data collection and analysis with the theories, constructs, and measures developed in academic disciplines related to government functions—whether they be in education, housing, transportation, health, environment, or energy. It was a hard sell and, finally, two of our eight panelists declined to be parties to this advice.

In truth, NCES has been diligent about making these sorts of connections. The message of the authors is that those efforts must continue. One lesson here for NCES is to work out an understanding about what the relative contributions of research and statistics can be, then to call on other OERI resources when that is appropriate and use the methodologies of both if that is called for. A dazzling example of the using both is described in Jim Stigler's paper "Large-Scale Surveys for the Study of Classroom Processes."

Another lesson for the Center is systematically to search out, use, and adapt research findings in its statistical activity. A fresh example here is the result of 5 years' work on school restructuring conducted by the University of Wisconsin, which was reported a few weeks ago at a seminar in the House of Representatives Rayburn Building. The researchers reported that, for the first time, their work showed a particular approach to teaching was associated with increased student performance—and that increase was for all students. The approach, which the Wisconsin Research and Development Center calls "authentic pedagogy," requires students to think, to develop in-depth understanding, and to apply academic learning to important, realistic problems. I cite this as just one example where the Center should, if these impressive sounding findings hold up, make use of research results as it designs data collections, questionnaires, supplementary information sources, and analyses on teaching.

ON TECHNOLOGY AND ADMINISTRATIVE RECORDS

One implication of the technology and administrative records concepts advanced in these commissioned papers is that the Center would actually collect less data. The emphasis here is on the word "collect." As I read what Fritz Scheuren and Glynn Ligon have said, and others have touched on these themes as well, the combination of electronic record keeping with administrative records is likely to result in much more information contained in those systems than has been the case with paper files. Thus, information that has, up to now, been obtained by surveys may in some cases be available with little investment of either time or funds. At a time when the government funding outlook is lean, while the data demands seem to grow, this combination appears to offer a vision of more for less that is pretty attractive under the circumstances. It does hold out the intriguing possibility that the Center might actually be able to achieve some of the other quality data, analytic, and methodological enhancements called for in this group of papers.

I have two caveats about such a development. First, while NCES has already worked hard on technology advances, it has not absorbed into its data planning and gathering how its own role may need modification in relation to that of states, colleges, and universities and other data providers and users. Perhaps furthest along in this regard are the Center's efforts on library data where self-editing electronic reporting has been developing over several years. Center staff would need to be strongly oriented to the needs of the institutions that are developing electronic systems but, at the same time, alert to how cross-state and national needs could be achieved as well. When and how to intervene, how to play the broker role, and when to subsidize design or planning efforts that can benefit many data needs are examples of the sorts of roles NCES would perform more frequently. Still, these roles are not the same as those exercised in designing another federal survey and may imply a need for staff development activities.

The second caveat is not to leap too quickly to a conclusion that all the public's data expectations for the Center can be derived from electronic systems with their low marginal costs. There will be issues or topics these systems simply do not cover that are a necessary part of an adequate education information system for the public. Examples include teacher practices and attitudes, student experiences, family circumstances, and attitudes that will and still only be available from individuals. Another example is that the present state-of-the-art in test equating will not permit ready comparison of student achievement results obtained from electronic records and based on widely differing assessment systems. And still another, if the appropriate data to answer a question must come from observations of classroom activity—as in the international classroom video research, which as David Mandel insists is essential when teacher quality is to be adequately measured—electronic administrative records will not contain those either. There will still be much for a statistical agency to do even in this wondrous new electronic world.

ON ANALYSIS

The final matter I want to address in these introductory observations has to do with implications in the commissioned papers that NCES would be more aggressive in making use of data in ways that will better inform the American public about education. This means more analysis. Among other things, it means more tapping of data from a variety of sources—OERI-sponsored research, the Bureau of the Census, systematic reports from organizations such as the Education Commission of the States or the Council of Chief State School Officers, and in-depth studies from individual states or districts, as well as from the Center's own collections. Here, too, the Center can expand its efforts with such organizations to design studies so that their data can be linked and, thereby, made much more powerful for analytic purposes.

Most challenging, however, will be the consideration of breadth versus depth trade-off questions. Reading through the full set of commissioned papers conveys both of these dimensions and could easily lead to numbing paralysis in formulating an appropriate Center response. Program and policy evaluations, longitudinal studies, data at the classroom level, experimental designs, international comparisons, and other costly steps are recommended by the authors.

What to do? The Center has the advantage of context—its place, function, and visibility—in picking the issues about which it will inform Americans. But the cumulative advice in these papers is that to inform well, the Center must narrow the questions it designs statistical activities to answer in favor of providing more powerful and complete data about the questions it chooses to address. The familiar dilemma in the Center is that of the longitudinal studies, such as High School and Beyond and the National Educational Longitudinal Study, where compromises must be made on education questions in relation to, say, employment or family background. Unfortunately, sometimes the result is that important policy issues and data relationships simply cannot be examined with these data. I do not mean to suggest there is an easy answer here, but increasingly the Center is being advised that what constituted adequate data in the past is frequently insufficient for the more sophisticated education information questions being posed as the Center prepares for the next century.

CONCLUSION

These are only a few observations that might be made, and I hope they will encourage readers to delve into the full volume of papers and commentary. The Center certainly received its money's worth from these papers. They are not full of impractical ideas. Instead, they provide the basis for a new vision of the National Center for Education Statistics in a new era.

2

Tracking Education Reform: Implications for Collecting National Data Through 2010

Tracking Education Reform: What Type of National Data Should Be Collected Through 2010?

John F. Jennings
Diane Stark

ABSTRACT

This paper will focus on the types of data collection the National Center for Education Statistics (NCES) could undertake that would be most useful to policymakers as they address issues of school reform. In particular, this paper will address three questions: 1) what type of education reform data should be collected; 2) how should the data be reported and packaged so that it is useful to policymakers; and 3) what should be done with the data so that policymakers have a better understanding of education reform?

We make three major recommendations to NCES for making its data collection efforts more useful to policymakers. First, we propose that NCES broadly collect data on specific education reform efforts being undertaken at the local, state, and national levels. However, understanding the cost implications of such a broad data collection, we urge NCES to augment its own efforts by compiling education reform data that has been collected by others, and to report both the NCES data and the compiled statistics. Second, we recommend that the education reform data be reported by state, and that information be available to policymakers via the Internet. Finally, we urge NCES to make the data widely available so that researchers and others will be able to undertake in-depth analyses, thereby enabling policymakers and others to have a better understanding of the reform being examined.

WHAT TYPE OF EDUCATION REFORM DATA SHOULD BE COLLECTED BY NCES?

For well over a decade, the nation has been concerned with improving education, especially at the elementary and secondary school levels. Therefore, in order to keep policymakers and the public adequately informed, there needs to be more information collected on education reform. In order to do this, NCES will first need to determine exactly what "education reform" is, and will then need to develop common definitions of the various reforms so that reporting will be uniform. Policymakers and the public will also want to know what effect the reforms have had on student achievement. In making determinations about student achievement, NCES will need to utilize the National Assessment of Educational Progress and other assessments and studies. Finally, NCES must report on both the statistical data as well as the student achievement information.

Defining Education Reform

Because education is an activity that falls mainly in the public domain, information on education is vital. Voters need information on their schools to determine whether or not to support increased funding for the school system. Similarly, parents need information to determine if the school is providing an adequate education for their children. Finally, educators and policymakers need information to make decisions regarding curricula and approaches to teaching. The National Center for Education Statistics (and its predecessor agencies) has carried out its mandate to report on the status and progress of education in the United States for nearly 130 years, providing policymakers and educators with a broad array of statistics and other data on our schools. As the Center endeavors to collect data on education reform, we would urge a similar broad-based approach.

At first, defining the term "education reform" appears to be uncomplicated because it could be defined simply as any effort undertaken at the local, state, or national level to improve student achievement. However, the task gets more difficult if one considers that what one person may view as a "reform," another may view as an impediment to improvement. For example, one policymaker may consider private school vouchers as an education reform, while another may view such vouchers as a step toward the destruction of public schools. Similarly, a group of policymakers may advocate opportunity-to-learn standards as an education reform, while others see those standards as requiring unnecessary expenditures in education. In collecting data on education reform, care must be taken to ensure that the Center's data collection efforts in the area of education reform are not viewed as politically motivated or as biased toward one set of reforms.

We believe that only by employing the broadest definition possible of "education reform" and then collecting and compiling data on all reforms can NCES remain an impartial statistical agency. In order to ensure its impartiality, the Center may want to consider convening an advisory committee, made up of individuals with widely varying views on education reform, to guide it as it embarks on this work. Such a committee would not only help the Center determine the broad array of education reforms to study, but also, given the reality of funding constraints, could aid in establishing priorities for NCES's data gathering in this area.

Once the Center determines which education reforms to study, we propose that NCES collect and compile a broad array of data that could be issued in a comprehensive annual report. Such a report would be a vital source of information to policymakers: by having access to information on what other communities or states are doing to improve education, policymakers would be able to "borrow" ideas and secure expertise that would help them in their efforts to improve education in their communities. To our knowledge, no such "encyclopedia of education reform" exists.

NCES will need to broaden its current statistics and data gathering efforts in order to report on specific education reforms. The advisory committee convened to identify the broad range of reforms could also develop common definitions of terms. While many states and communities appear to be undertaking similar education reforms, such as charter schools or school choice, there are likely to be great and small variances among them. Common definitions will allow for uniform, comparable reporting of data. Once common definitions are developed, NCES should then direct its efforts to collect data that will paint a complete picture of the types of reforms being undertaken, and that will enable researchers and others to undertake analyses that will help to determine which reforms have been successful in improving student achievement and other outcomes.

Many NCES ongoing data collection efforts, studies, and assessments will provide additional information that is needed to determine the effects of education reform. For example, through the National Assessment of Educational Progress (NAEP) and the various NCES longitudinal studies, the Center is able to provide data on student achievement as well as in-depth information about students and the education they receive. This information, along with other data collected by the Center, such as dropout rates and student course-taking patterns, will be helpful in giving policymakers and others the information they need regarding education.

If NCES were to develop a plan for gathering this ambitious set of data on education reform, perhaps there would be interest in the Congress to devote additional funding for the plan's implementation. However, given the current fiscal climate on Capitol Hill, NCES may want to consider other means of financing the data collection. Perhaps the Center could establish a "fund for education reform statistics" and work with the business community and charitable foundations to contribute additional dollars so that adequate funds would be available to carry out this work.

If the Congress does not provide extra appropriations for education reform data collection, and if NCES opts not to create a special fund, then the Center must set priorities for which information it intends to collect directly (perhaps following the advice of the advisory committee). As a means of lessening the burden and allowing for the reporting of a broad scope of education reform data, NCES should also consider becoming a "repository" of data collected by other organizations. Many organizations such as the Council of Chief State School Officers (CCSSO), the Education Commission of the States (ECS), the National Conference of State Legislatures (NCSL), and the American Federation of Teachers (AFT) have been collecting data on various state education reforms. NCES need not "reinvent the wheel"; the information gathered by these organizations could be included in any reports on education reform. We understand that the compiling of data from other sources may raise questions of validity and reliability; but with limited funds available to NCES for a comprehensive education reform data collection effort, it may be the only alternative. NCES could work with CCSSO, ECS, NCSL, AFT, individual states, and private research groups to use common definitions and reporting cycles so that this data could be included in a comprehensive annual NCES report on school reform.

We also recommend that NCES go "on-line" with a page on the Internet devoted entirely to education reform. NCES's education reform "home page" could provide a monthly update of the data that appeared in the annual report on education reform and could also include recent research findings or data from outside groups. For example, there could be a section on charter schools that would include any updated information that may be issued by the Humphrey Institute in Minnesota. Similarly, any data on state standards-based reform efforts that are issued by NCSL, ECS, AFT, or other groups could be included in the section on standards-based reform.

To assist NCES in thinking about the wide range of school reforms that could be included in its reports, the following sections will briefly describe several major types of reforms.

Standards-Based Education Reform

One of the most prevalent reforms currently being undertaken by states and school systems is standards-based education reform. The underlying premise of this type of education reform is that students, teachers, and parents should know in advance what students are expected to know and be able to do at different grade levels. States develop content standards outlining what is expected in subject matters such as math and science. States also establish performance standards that explain how well a student should perform in a given curricular area, and develop assessments aligned with the state content standards in order to chart student achievement. According to a recent report issued by the American Federation of Teachers entitled *Making Standards Matter*, as of July 1995, 49 states and the District of Columbia were engaged in standards-based education reform. The federal government, through the Goals 2000: Educate America Act and the Title I program, is assisting states and localities as they develop standards and implement this comprehensive reform.

Several national organizations have issued reports on various aspects of state actions in standards-based reform, and other offices in the U.S. Department of Education, such as the Goals 2000 office, have also compiled information in this area. NCES could use the information compiled by these organizations and by other offices in the Department as a basis for issuing a comprehensive up-to-date report on standards-based reform. It would be of great value to national and state policymakers to have a regularly updated status report of which states are developing, have adopted, or are implementing standards in a given subject matter area. A policymaker from a state with no science standards, for example, could use this information to seek out officials from other states that have adopted science standards, and could consult with them about their standards and the process by which the standards were developed. Such a status report would also help state officials to determine where they "measure up" compared to other states implementing standards-based reform. National policymakers would also find the state-by-state standards status report useful as they make decisions regarding the Goals 2000 program and the Title I program.

An integral part of standards-based reform is assessing student achievement. Most states are just beginning to develop assessments based on their standards, and there is little information available to states to help them during this process. Data should be collected on where states are in the process of developing assessments that are aligned with standards, and the type of assessments being developed by the states.

Finally, a controversial component of standards-based reform that some consider key are opportunity-to-learn standards. These standards outline what tangible elements need to exist in a school in order to give a student an opportunity to learn the state standards. They are controversial because they primarily affect "inputs" into the school system (thereby potentially requiring an outlay of funds), whereas the content and performance standards are concerned only with outcomes. For example, if a state science standard requires that by the 5th grade students should know how to operate a telescope, the opportunity-to-learn standards could require that elementary schools have telescopes. NCES should collect data on the number of states that have developed or implemented opportunity-to-learn standards, and where possible, determine if the implementation of these standards has had a fiscal impact.

Other Reforms

There are several other reforms that are being tried around the country that may affect only individual schools rather than the entire school system, as is envisioned in standards-based education reform. These reforms are either being implemented or considered in nearly every state in the nation and should be examined since they are so pervasive. State-by-state information on the number of schools implementing a given reform would be helpful to policymakers, especially at the national level where some federal programs have been created to encourage certain types of reform such as charter schools. These reforms should be followed over time in order to develop trend data. The following sections outline some of the education reforms that states and communities are undertaking that NCES may want to study.

State Course-Taking Requirements

During the 1980s, after the issuance of *A Nation At Risk*, many states changed state curriculum requirements. Under this reform, students were required to complete more hours of instruction in core academic subjects in order to receive a high school diploma. NCES has done much work in this area through the High School Transcript Studies and should include this updated information in any comprehensive reports on education reform.

Charter Schools

Charter schools are public schools that are by state law exempt from significant state and local requirements. In exchange for this increased flexibility, the schools are held accountable for increased student achievement. It is believed that by exempting these public schools from most rules and regulations, charter schools create an environment where innovation can thrive. They are often created by teachers, parents, or groups in the community. Basic data should be gathered on the number of charter schools in the United States, and because charter schools exemplify the trend in public education of "greater flexibility in exchange for greater accountability," they should be studied over time to determine how successful the charter school approach is in raising student achievement.

Reform Networks

Hundreds of schools across the nation are engaging in reform activities that employ special strategies for educating children, especially those from disadvantaged backgrounds. Examples of these reforms include Robert Slavin's Success for All, Hank Levin's Accelerated Schools, and Ted Sizer's Coalition of Essential Schools. In an effort to provide support for schools undertaking these above reforms, networks are created where individual schools can go to get the assistance they need. Data on the number of schools participating in the reform networks by state and, where possible, by school district, should be included in any education reform report.

School Choice

School choice programs allow parents and students to determine which schools to attend. Many states and school districts have choice programs that allow students, within certain restrictions, to enroll in the public school of their choosing; some states also have choice programs that allow students to attend private schools using public funds. Other forms of school choice would include magnet schools created to promote integration of different racial groups. Information needs to be gathered on the number of states and school districts with school choice programs; the number of students participating in the choice program; and examples of state or school district choice policies (i.e., open enrollment, student selection criteria, and so on).

Private, For-Profit Companies

Several school systems have contracted with private, for-profit companies to run their school systems or to operate individual schools. Data should be gathered on the number of school systems that have made such arrangements, which companies have been involved, and general characteristics of the schools or school systems that are affected.

State Takeovers

While not quite an education reform in the traditional sense, in an effort to boost student achievement, some states have taken over failing school systems. Data should be gathered on the number of states that have a policy or law allowing state takeovers; the "triggering" conditions for such state intervention; and the number of school districts or schools that are affected by state takeovers.

School Finance

As part of any report on education reform, data should be included on state and local efforts regarding the financing of education. It appears that it is a trend among several states to limit or curtail the public financing of education through local property taxes and to instead fund schools through state taxes. Meanwhile, there are several court cases pending that call into question the disparate per-pupil expenditures existing within states. Further, there is some concern about differing per-pupil expenditures within school districts. Data should be reported, on a state-by-state basis, on all aspects of the school finance issue. On the national level, if the dramatic funding cuts that have been proposed for federal elementary and secondary education programs become law, it would be essential to know how these reductions affect school finance at the state and local levels, including the number of students who no longer have access to the programs supported with federal funds.

Block Grants

There are various proposals before the Congress to create education block grants. Block grants are formed by taking a number of separate programs that have similar purposes and combining them into one large program with one set of requirements. Block grants usually mean increased flexibility in the use of federal funds by states and school districts, but the creation of a block grant is also usually accompanied by decreased federal funding. If such education block grants are enacted into law, they will have a considerable effect on the financing of education, and should be studied to determine their impact. Several states are also "block-granting" state categorical programs, and these efforts should be reported.

School Infrastructure

Another element of school reform that is not instructional-based is the renovation or rebuilding of aging school buildings. As illustrated in Jonathan Kozol's *Savage Inequalities*, some school buildings are in such poor condition that they are literally falling apart, lack working plumbing, and are unfit for human occupancy. In other instances, the buildings lack the necessary facilities (such as science laboratories) to adequately provide the type of education that is needed in today's high-tech world. Data on aging school buildings and the state of school facilities need to be included in any reports on education reform so that state and national policymakers will have information about the general condition of schools across the nation as they make policy decisions.

School-to-Career Reforms

Through its Data on Vocational Education (DOVE) system and the first and second National Assessment of Vocational Education (NAVE) reports, NCES and the Office of Educational Research and Improvement have gathered and reported significant data on the status of vocational education in the United States. Continuing these data collection efforts is extremely important as reforms in vocational education, such as "tech-prep" programs and school-to-work transition initiatives, grow in popularity among the states, and as these initiatives become a primary instrument for reforming secondary schools.

Home Schools

Home schooling is a growing trend resulting from parental desire to oversee all aspects of a child's education. Reasons cited for home schooling range from religious beliefs to dissatisfaction with the education provided by traditional schooling. It is important that data be collected on home schooling in order to get a clear picture of all the endeavors being undertaken to educate children. Again, state-by-state data on home schools are essential, as well as a description of the oversight governance of home schools in each state or locality. (That is, does the state, school district, or other entity ensure, through assessments or other means, that home-schooled children are receiving an adequate education?)

Postsecondary Education Reform

Nearly all the above-mentioned reforms affect only elementary and secondary education. Several states are beginning to examine their postsecondary education systems and are considering implementing reforms to improve teaching in higher education. Because this is an emerging reform, data on the number of states that have postsecondary education reform initiatives as well as information on the content of these initiatives would be very useful to policymakers at all levels, especially as more states embark on higher education reform.

Assessing the Goal of Reform

Elementary and Secondary Student Achievement

In studying all these reforms, NCES should gather data on what is the intended outcome of the reform. For example, nearly every reform mentioned above would probably have as one of its goals increased student achievement. NCES, through its traditional measures such as the National Assessment of Educational Progress (NAEP) and through other surveys and studies, could measure the impact of various reforms on student achievement. In particular, as NAEP becomes more aligned with the national academic standards, it can be an important vehicle for measuring student achievement under standards-based education reform. Other measures, such as the National Education Longitudinal Study and the High School and Beyond reports should also be used to determine if various reforms have improved student achievement. Additionally, NCES should also consider other sources of information on student achievement, such as state assessments and college entrance examinations. Each of these assessments is potentially a rich source of information on student achievement along with information on student characteristics.

Postsecondary Student Achievement

NCES should also look into measuring the achievement of postsecondary students. As was mentioned above, attention is beginning to turn to postsecondary education reform, with talk of developing academic standards similar to those developed for elementary and secondary education. Currently, NCES collects data on such factors as postsecondary enrollments, completion rates, the number of students receiving financial aid, and faculty and institutional characteristics. The Center does not assess postsecondary student achievement. While current budgetary constraints may preclude it, we would suggest that NCES consider either expanding NAEP to include students in postsecondary education or to develop a separate NAEP-like assessment to chart the achievement of students who have continued their education beyond high school.

Other Goals

NCES should also collect data that will enable policymakers and the public to determine if other goals of the education reform initiatives are being achieved. For example, while increased student achievement may be a goal of a tech-prep program or a school-to-work transition program, those reforms also have as their goal developing student occupational and work skills. Similarly, school choice programs may not have increased student achievement as their primary outcome; rather such initiatives may be solely designed to give parents more educational options for their children. These other goals of education reform should not be overlooked by NCES.

International Comparisons

In recent years, there has been increased demand among policymakers and educators for information on how student academic performance in the United States compares with students in other nations. NCES, through the OECD International Education Indicators Project and other international studies, has helped to provide that needed information. In the future, individual states will want to know how their students compare to students in other nations. NCES should be prepared to provide such comparisons.

PACKAGING AND REPORTING OF DATA

In order to be most effective in reaching policymakers, NCES needs to pay particular attention not only to the scope of data collected but also to its packaging and reporting. On all levels of government, policymakers' attention is being drawn in several different directions on many issues. Therefore, concise, timely information that is relevant and easily accessible is essential.

Readability and Relevance

Information for policymakers should be in a form that is easily understood since they are often dealing with several divergent issues at once and may not be experts in education. Reports that contain executive summaries as well as charts and graphs with easily understood explanations best suit the needs of policymakers. This sort of concise information would also be of use to a wider audience such as parents, the public, and the media.

Further, in order to meet the needs of policymakers, reports issued by NCES should contain data that is regionally or locally relevant. We recommend that NCES consider studying education reform on a state-by-state basis, and issuing annual reports on state activities. The state-by-state information would assist policymakers at all levels to understand the impact of education reform. More detailed information could be put "on-line" and be available to state legislators and others who may need more in-depth data on the condition of education reform within an individual state. Also, NCES may want to create a special state education reform hotline so that policymakers and others could have immediate access to the information.

In compiling the state-by-state data, NCES should, to the extent possible, collect data on education reform at the sub-state level, especially at the school district level. While we understand the cost and data reliability issues involved in sub-state reporting, we believe that this information is essential in helping parents and local communities understand what their schools are doing and if the reform being implemented is effective. This, in turn, helps policymakers do a better job at representing their constituents' views on education, as well as provides them with information they need to make informed decisions. If the Center is able to collect data on school district education reform, these data should include information on the demographic and economic makeup of school districts so as to give the reader a context for the information. Also, as a means of reducing costs, NCES could draw on state assessment data and other state reports in its sub-state data gathering efforts.

Timeliness

One of the ways that NCES can be most effective in meeting the needs of policymakers is to anticipate when certain data will be needed. This is especially true in the legislative arena. For instance, if the Congress is debating a bill designed to reduce student dropout rates, a report filled with state and local dropout data and an analysis of state and local dropout prevention programs should be issued before that the debate occurs so that informed decisions can be made. Usually, federal education programs are authorized for a set number of years, and near the end of that authorization period, the Congress begins to consider the effectiveness of the program and whether it should be continued. To the extent possible, NCES should pay close attention to the reauthorization schedule, and time the issuance of reports to coincide with that schedule.

Also, it goes without saying that NCES should provide up-to-date information whenever possible. Policymakers need to have the most current information available so that they can make decisions based on what is happening in the present, not on 5-year-old data. By updating the education reform data that would appear on NCES's "home page" on the Internet, policymakers and others would have immediate access to the most recent information available on education reform.

ANALYSIS

The most important aspect of making NCES education reform data useful to policymakers is to have more analysis of the data. While policymakers find it useful to know the postsecondary attendance rates of students in their school district or to know how well their students did on the state NAEP assessment, they are more interested in knowing why a certain achievement trend is occurring. In education reform, it is essential to know possible reasons why one reform succeeded in a state while another one failed so that policymakers can fine-tune programs or make other necessary adjustments. This information can be gleaned only from analysis of data that have been collected over time.

While NCES does engage in some data analysis, far more needs to be done, especially if the Center begins an effort to study education reform. NCES may not have the capacity at this point to conduct the kind of analysis that is necessary, nor may it be an appropriate function of the Center. However, it is our hope that the information collected and compiled by NCES on education reform and student achievement will spark the interest of researchers to look deeper into the data. In order to move that process along, we recommend that NCES, together with the Office of Educational Research and Improvement (OERI), convene a conference to discuss how NCES and OERI can make data more accessible and promote analysis.

To the extent that funding is available, NCES may also want to explore the option of contracting for such analysis. If NCES decides to promote analysis of education reform data through contracts, it should do so with several target audiences in mind. For the needs of policymakers (and for that matter, the media as well as the general public), the analysis and reporting should concisely explain the effect of an education reform and give possible reasons why this effect occurred. For the needs of educators and researchers, the analysis should be more complex, painting a more complete picture of how an education reform affected student achievement. Policymakers may also return to the more in-depth analysis as situations warrant.

SUMMARY

In conclusion, in order to meet the needs of policymakers, we recommend that NCES expand its data-gathering efforts to include reporting on specific education reform initiatives being implemented at the local, state, and national levels. Once the data are gathered, NCES should package the information in an annual report that would include education reform data collected by other organizations. The information contained in the annual report should also be available through the Internet and be updated as newer data become available. Summary information should be provided to policymakers, and NCES should try to issue the release of information in a manner that provides policymakers with timely, needed information as they begin to debate an issue. Finally, NCES needs to encourage more analysis of education reform data by researchers so that policymakers and others will be better able to understand possible reasons why a certain education reform succeeded in improving student achievement or why it failed.

This paper has outlined a rather ambitious set of recommendations for NCES with regard to policymakers' need for education reform data and statistics. We realize that meeting all of our recommendations may be impossible, especially given today's funding realities on Capitol Hill. However, we believed that we had to set out a broad vision so that actions could be taken to achieve it, if only partially. NCES must be sensitive to the changing needs of policymakers not only to be able to serve them better but also to remain a viable federal agency.

Where Are We Going? Policy Implications for Data Collection Through 2010

Christopher T. Cross
Amy Rukea Stempel

INTRODUCTION

We know that education reform is happening and that the academic achievement of American students is lagging behind what is expected of them both in our nation and in the world. However, we do not know what links education reform efforts to changes in academic achievement. The collection of educational statistics by the National Center for Education Statistics (NCES) can assist attempts to pinpoint the relationship between reform and achievement. By isolating different aspects of education reform and attempting to remove superfluous influences, we can begin to cull out those changes that are transforming American schools from those that are not.

What is rapidly becoming apparent is that even though we have some idea as to what affects academic achievement, we are still floundering in our efforts to reform the nation's education system. There seems to be no one reform that actually accomplishes all we need it to do, as much as proponents of various reform agendas might wish it to be so. However, there may be clusters of reforms that, when integrated and advocated with intelligence and moderation, might actually produce results. Unfortunately, we have little data to support reform recommendations of this type.

Besides the statistically reported national academic achievement and the change in that achievement over time, the Council for Basic Education (CBE) would like to suggest collecting data on other, less immediately apparent, factors in education reform. We do not discount the necessity of gathering statistical information on achievement; however, we believe that our biggest pitfall in education has been ignoring the more subtle issues affecting reform.

Beyond data collection, NCES might consider devoting more attention to analysis of that data. If in-depth analysis is not feasible, providing readers with possible considerations for analysis would be helpful. There is no doubt that the efforts of NCES to encourage wider use of their data have met with success. CBE would like to see this effort continue to be a NCES priority. The challenge of statistical analysis is to get to the heart of why and how education reform helps or hinders educational achievement and student learning. To do this, we believe that the questions asked will have to be modified to capture the inherent ambiguity and interconnectedness of the educational endeavor.

One of our major concerns is the validity of the data collected. Over the years, we have discovered that self-reported data are notoriously unreliable. For example, teachers over-report their implementation of reform; principals assure the public that their schools are consistently performing better now that "X" and "Y" reforms have been mandated; and parents seldom understand reform agendas enough to make informed decisions about the truth or falsehood of these statements. Without reliable information, we run the risk of making ill-informed decisions that will do more harm than good. By ensuring that the data collected are valid and by encouraging analysis, if not providing it, the information collected can be put to direct use by policymakers, educators, parents, and students.

VALUE ADDED

Information on achievement is a "slippery fish to catch." We have learned over the years that there are many factors that affect student learning that are not school functions. For example, countless statistical surveys have shown that socioeconomic status, specifically the mother's educational level, is more of an indicator of student achievement than any other factor. With all these secondary indicators of success floating around, we need to be careful to what we ascribe achievement and how we report it.

Therefore, we believe strongly in the need to examine the value-added issue more closely. Given a variety of starting points, what does a particular reform effort or combination of reform efforts add (or not) to current student achievement in a particular school or type of school? If we can begin to address this question, we will be well on our way to establishing the relative merits of various reforms. Data collection in this instance should focus on the school, the types of interventions offered, and how they affect a variety of students attending. While other factors may be stronger indicators of success, schools are ultimately more manipulable than an individual's SES or parental education levels.

There are many reasons to worry about education in America. Our highest achieving students are lagging behind their world counterparts. Even more debilitating are our inner city schools, many of which seem barely able to teach students to read, let alone succeed in the world. Scores on achievement tests alone may or may not validate the success or failure of reform efforts for both the highest and lowest achieving students. However, looking closely at the changes in achievement scores and their relationship to many indicators, rather than accepting them at face value, would come closer to answering the question: What is the value added by this school? this reform? this program? Fundamentally, we believe that this is what people want to know.

We have developed our recommendations with our need for reliable information and our belief in the need to articulate and ferret out the value added by various reforms in mind. Education is complex and educational data collection needs to reflect this. Unfortunately, this presents a different set of implications to those collecting data. The point of this paper, as we understand it, is not to support the status quo, but to go beyond the traditional role of NCES and challenge it to find solutions to the intricacies of reliable, complete, and insightful educational data collection.

The reform agendas discussed in this paper are based on an assessment of what is present on the national reform horizon. Subsequently, we will be discussing teacher education and development; issues of school governance and organization, such as site-based management, home schooling, charter schools, magnet schools, and voucher programs; articulation between levels of schooling; educational technology; academic standards; and assessment. As you will note, there are particular dimensions of these issues in which we are primarily interested that will constitute the bulk of this paper. We will also discuss the costs and benefits of some of our suggestions, followed by a brief discussion of the implications these suggestions have on data collection methods.

THE DIMENSIONS OF REFORM

Systemic information about education reform efforts is crucial because the long-term health of our national education system requires that we radically change how we educate our students. How does our education system work? What are its flaws? Why is change so difficult? By examining and collecting data on crucial systemic points, we can begin to address the complexity and interconnectedness of education reform. We are primarily interested in how these reforms are being used (or not) in the system. However, more important is what effect these reforms have on academic achievement. Throughout this paper, we will consistently return to these two issues.

Teacher education has been under siege for many years. Tomes have been written on what teachers are not required to do to become certified and how undertrained and ill-used they are. CBE suggests an attempt to evaluate and measure the quality of teacher training, professional development, and professional support in an integrated way. Rather than providing a catalogue of courses required to become a teacher, it would be more informative to provide information about the philosophies of particular schools of education and how these philosophies are carried out both in the teacher training curriculum and later in the K–12 curriculum.

For example, how does the curriculum of teacher training institutions that advocate student-centered learning reflect that philosophy? How is a vision of the educational experience linked to its practice? How do the professors responsible for training future teachers conduct their classes—lecture, group activities, socratic seminar? How do teachers trained in a particular institution translate their training into the classroom? Discovering what actually goes on in teacher training programs and classrooms, rather than what is reported to happen, would help us make decisions as to what works and what does not.

Another reform agenda, "site-based management," has recently been coming under fire from the public and policymakers. We would like to determine the extent to which site-based management exists and is working nationwide. How is site-based management defined? In schools where it is said to exist, how is it implemented, and are the results substantially different in practice from the norm of top-down management? What is gained and what is lost by switching to site-based management? Are there similarities in site-based management and the administration of private schools that are worth exploring? And last but not least, does successfully implemented site-based management have a positive effect on student learning?

Given that the purposes of schooling are to prepare students for a personally and professionally productive life, more consideration needs to be given to how the system works together (or not) to support a consistent purpose and vision of education. One way to begin this, CBE believes, is to develop a national survey that examines the articulation between the levels of schooling. Anecdotal information indicates that it is surprisingly uncommon for teachers or principals to examine what happens in the levels of schooling before or after the level that is their immediate responsibility—whether it be 7th grade specifically (teachers) or middle school in general (principals). Similarly, we suspect it is uncommon for teachers or principals to examine what teachers of other disciplines at the same level do and expect in their classroom. We would like to know if there is horizontal and vertical coordination of curriculum, and what are the expectations within individual schools and throughout school districts. For example, what are the expectations of the elementary schools in a district? Are the middle schools in the district aware of those expectations and do they begin where students have left off? Does a teacher in 8th grade, for example, know what other 8th-grade teachers do and expect in their classrooms?

There are other issues currently in the public interest that are not central to student instruction but that deserve exploration: for example, school safety, moral and character education, and the effects of parental involvement on students' educational achievement are "hot" issues in the reform debate today. What these proposed reforms have in common is that they do affect student achievement, but no one is sure how they interact with other elements of reform. Does a rigorous character education program improve academic achievement or school safety? Does simply making a school safe improve behavior or academic achievement? Is parental involvement in their children's education linked to the other factors discussed? An informative survey would try to tease out the different strands of reform and establish how they interact. No doubt this would be a complex data collection to design; but one thing is certain: it would be incredibly helpful to all involved in education.

Data needs to be collected about several more recent phenomena: technology, home schooling, charter schools, magnet schools, and voucher programs. How do these efforts at school governance and organization affect academic achievement? Technology offers us a new way to gather and disseminate information and provides an ease in data, word, and information processing previously unknown. Unfortunately, the education community, for lack of money and political power, is far behind the technological boom. In order to truly document how technology affects learning, we will need to document the uses and abuses of technology in classrooms across the country. For example, recent newspaper reports have revealed a frustration that technology is not the panacea it was first touted to be. We doubt that there is a single panacea, but even so when the reports were followed up, it was discovered that the computers were being used as high-tech workbooks and that the learning process had not significantly changed. Just because schools have access to technology does not mean they use it to its fullest capacity. How schools use the technology they have is one of the more crucial questions of the next few years.

A related issue is the state of the technological infrastructure of the nation's public schools, which is dismal. No one really knows how much money it would take to upgrade them. The true extent of the problem is often obscured by the massive amount of speculation and little hard data. Often surveys ask if a school uses computers in its classrooms, but not how many modems or Internet connections each classroom has. As the uses of technology grow, so should our interest in how schools are putting their technology to use.

As with the technology issue, home schooling in the United States is rapidly increasing. However, we lack ready information as to who does it, why they do it, the average number of years they do it, the most popular grades to home school, and how technology has affected the home-schooling boom. We also lack information about the academic achievement of these students and how they fare when (or if) they return to school. Where home schooling works well are there lessons that can be transferred to schools about use of time, student/teacher ratios, and options for creativity? It would also be important to find out whether home schooling is a growing option or a passing fad. Although home schooling is often ignored, we believe this segment of education is one of the fastest growing, with research potential yet untapped.

Charter schools, or public schools that are given permission to ignore certain rules and regulations in order to try to increase student achievement, are also growing by leaps and bounds. Because they are free of crippling bureaucracy, charter schools have the flexibility to implement reform decisively. However, for every success story, there are instances of financial abuse and declining test scores. We suggest attempting to evaluate the performance of charter schools nationwide. One fundamental tension in education addressed by charter schools is accountability versus flexibility: what is gained and what is lost when schools, administrators, and teachers are given autonomy in their decision making?

Magnet schools are public schools that have been allowed to choose a particular focus for their academic activities. Perhaps the most famous of these is the public High School for the Performing Arts in Manhattan, which was the inspiration for the movie *Fame*. Aside from the performing arts, schools can be organized around marine studies, the military, technology, or even traditional pedagogy, to name a few. Students must choose, and be chosen by, the schools. The interesting elements of magnet schools are the focus provided by the organizing principle and the element of choice. Does a thematic approach to an academic education provide a focus and motivation for students? Does the fact that students must choose the school and make a commitment to it, as well as be chosen by the school, increase academic achievement?

The increase in voucher programs throughout the country also deserves consideration. School vouchers provide parents with a certain amount of money per student, commensurate with per-pupil expenditures in that district, which they can take to either public or private schools within their community. Is there an increase in academic achievement for students whose parents use vouchers to choose schools? How much of an increase or decrease in achievement is due to the particular school and how much is due to the act of making a choice and a commitment? While difficult questions to answer, the results of such an inquiry would enable parents and policymakers to make decisions about school choice more effectively.

Besides systemic information, there is also information about what occurs in the classroom that is crucial to educators and policymakers. While systemic reform is necessary and desirable, the work of education goes on in the classroom, and it is there that we must look for the bulk of our information. Given that radical systemic reform is still a long way off, a detailed look at what is currently happening in classrooms nationwide will help us in our more immediate future.

The current debate about national standards is an interesting one at the policy level; however, there is no information about the implications of the standards setting projects on education reform. Do academic standards improve achievement? How are academic standards being implemented—top down or bottom up? In states that have developed academic standards, is curriculum being designed with standards in mind? Is professional development provided for those teachers who are expected to implement standards?

We suspect that most who support standards honor them in the breach; the inertia of schools tends to slowly make its way back to the status quo after a vigorous attempt at change. Effective data collection and analysis in all areas of education reform will enable policymakers, educators, teachers, and students to take the pulse of the system and measure what their responsibilities are in order for true reform to occur.

To even begin to measure the effect of academic standards in the schools, we will have to measure the quality and uniformity of assessments and testing in the classroom. CBE has discovered, in the process of doing business, that there are incredible assumptions made by parents, educators, and policymakers about the verity and uniformity of individual student grades. Therefore, we also believe that it is necessary to have some data collected about individual teacher grading schemes such as how teachers determine individual student grades, how they construct their own assessments, and what they fundamentally want their students to know and be able to do. Are teacher grading schemes uniform? This is information the public needs to know.

COST-BENEFIT INFORMATION

Hand in hand with information on instruction and learning, we also need to analyze the costs and benefits of various reform efforts. Reforms that might at first appear to be expensive prove to be quite cost effective when examined from the point of view of the benefits they will provide in teacher training, reduced need for remediation, and student focus, for example. Alternatively, reforms that at first seem inexpensive might prove to cost schools more money if they are not well-organized and meaningful.

For example, what are the initial costs of infusing technology into the schools? What are the maintenance and upgrade costs of educational technology? What services can ports to the Internet provide students and teachers that might take financial and time pressure off school districts—teacher training and development and access to archival records, for example?

Most interesting to reform efforts in the days of shrinking budgets is how successful schools streamline the use of limited resources. Perhaps the most useful question would be how do these schools set their instructional, hence financial, priorities and what are these priorities? Successful schools often employ creative methods to develop resources they believe necessary to instruction and learning. What are these creative methods? How well do they work? Are they personality dependent?

DATA COLLECTION

We understand our role in this process to be that of provocateur; hence, some of our suggestions of areas to explore will require different methods of data collection than those used in the past. We suspect that data collection efforts would have to become more delicate, sensitive, and focused. As always, there are advantages and disadvantages to such a change. By isolating very specific information, there is a limit to the number of ways it can be used. However, if we can determine that the data collected, no matter how focused, are able to provide pertinent information about the reform movement, then the trade-offs would be worth it.

Statistical data collection is certainly useful. However, CBE believes that its use is limited in fields like education where success or failure depends on a host of often conflicting variables and human imperfections for which there exist limited methods to control. We suggest using pure statistics as a tool for analysis, not as an end in themselves. In other words, provide people with the initial information to investigate "why" and encourage them to do so.

One way to do this is through an integrated combination of quantitative and qualitative information working together to answer questions about education reform. Because self-reporting is unreliable, we need to consider other options for data collection such as independent data collection agents or "inspectors" who are responsible for evaluating the relative levels of existence of various reform efforts. We do not let students grade themselves, so it seems equally self-defeating to let those who participate in schools be the sole assessors of their own success or failure. While perhaps blurring the line between data collection and research, we currently see no other way to ensure accurate, reliable information about systemic activity.

These data collections could take the form of both longitudinal studies and single-point studies. For example, in the teacher training example for data collection mentioned earlier, longitudinal studies would be most effective to determine the influence of teacher training on future classrooms, while a single-point study could help determine the context of teacher training.

We also recommend stair-step surveys to link information at the school, district, state, and national levels to what is going on in individual classrooms. Surveys of this sort would inform us as to whether the coordinated reform effort is going well or not.

We understand that we may be recommending an extension or redefinition of some of the activities of NCES. Please view our recommendations in light of our mission and position in the field of education reform. After 39 years of advocating rigorous liberal arts education for all students K–12, we believe that instruction and learning are not simple processes to be easily understood and broken down. As valuable as NCES statistical data are, we believe they can be made more valuable by extending their purpose and offering users even more reliable information, more subtly realized.

Discussant Comments

MARY J. FRASE

These two papers are very similar, and there is considerable overlap in their perspectives and some duplication in specific recommendations. Both represent the perspectives of policymakers and try to outline what the authors feel would be useful information about reform for policymakers. A major difference between the papers is the emphasis in the Jennings and Stempel paper on the need for state-level and sub-state data. Because of the similarity and overlap between the two papers, I will discuss them together, rather than each one separately.

Topics Related to Reform

Out of the two papers one can assemble a long list, a wish list, of reform-related topics one or both mention as being useful to have information about. The former Commissioner, Emerson Elliott, mentioned that planning efforts during his tenure tended to produce similar results—long lists of topics people wanted information about but no suggestions about what to delete. Out of these papers, I came up with 35 separate issues or topics (see Appendix A to this paper), and I probably missed some. In this case, the long list was probably deliberate in light of the charge to the authors not to take fiscal constraints into account. Jennings and Stark wrote that they purposely took a broad view, realizing that NCES probably could not do everything they mentioned.

One issue that needs to be raised at the beginning is what is meant by "reform." Is it just another word for "change"? The Jennings and Stark paper warns that NCES should take a very broad view of what constitutes reform so the agency would continue to be seen as impartial, i.e., not endorsing one approach or type of reform over another. But does that lead to looking at everything? Is there a trade-off here between depth of information and impartiality? If NCES were to gather a great deal of information on few, high-priority "reforms," would the agency be seen as endorsing those reforms? If it gathers information on a wide range of reforms, the result may be breadth but not depth of information. Jennings and Stark suggest an advisory panel to help set priorities about which reforms to follow. In the current fiscal situation, NCES cannot expect more money, so someone would have to make some choices. Perhaps the best way to think of these papers is to view them as a menu from which NCES could choose topics.

Toward that end I have tried to think not so much about specific "reforms," which may have relatively short "half-lives," but rather types of information about reform that might be useful to have, regardless of what the reforms are. I grouped the topics the two sets of authors mentioned into a limited number of categories in order to see if that might lead to some insights about what NCES might pursue out of this menu. Most of the 35 seem to represent one of four types of information.

• *How much reform is happening?*	How many students are being home schooled? How many school districts have school choice plans allowing choice within or across districts? How many states are pursuing standards-based reform?
• *How is reform being carried out?*	What oversight mechanisms are in place for home schooling? What types of school choice plans are being used? What does "standards-based reform" mean in the various states?
• *What is the effect of reform on achievement and other outcomes?*	How well do home-schooled children perform relative to those in public or private schools? How is the availability and the utilization of school choice related to student achievement, motivation, and parental satisfaction/involvement?
• *What basic kinds of data are needed to provide contextual or baseline information for reform efforts?*	How can NCES provide contextual information about school finance, student mobility, school facilities, postsecondary achievement, and teacher development?

I will briefly discuss each of these four types of information, as well as their relationship to the NCES data collection program. This is not the only way nor necessarily the best way to group these topics, but the basic point is the need to think systematically about categories of information rather than about specific, relatively narrow topics or issues (i.e., not to miss the forest for the trees). What types of information are most appropriate for NCES to gather directly and what role, if any, might NCES play relative to other kinds of information?

How Much Reform Is Happening?

This is the simplest of the categories. It involves tracking the extent of reform activities by collecting counts of different activities, once decisions are made about what to count. Such information can be collected with fairly simple questionnaires. While NCES can do some of this, it may not be a good use of NCES resources. Other parts of the Office of Educational Research and Improvement (OERI) and the Department of Education are already doing some of this, as are other organizations such as the Education Commission of the States. The regional laboratories are possible candidates, and Emerson Elliott mentioned that there is already some interest among the labs for doing something of this sort relative to charter schools. Charter schools are one of those relatively rare phenomena that NCES is not good at capturing.[1]

The role of NCES relative to this type of information might take two forms. NCES could collect it directly where there are existing vehicles for doing so, which might include the Common Core of Data (CCD), Schools and Staffing Survey (SASS), or Fast Response Survey System (FRSS). Alternatively, NCES could play a brokering role, where the agency would determine what holes existed in terms of missing information and would help identify other ways to gather such data, perhaps involving the Forum and the Cooperative Systems.

How Is Reform Being Carried Out?

Some of this also involves collecting counts of activities, but at a more detailed level. What kind of approach to gathering the information is most appropriate depends on the level one is interested in, i.e., state, school district, school, or classroom. Some kinds of information could be gathered with more detailed questionnaires. Others may require case studies of how reforms are being implemented, since the research literature on implementation reveals there is much slippage between written policy and what happens in the field, in this case, the classroom. The same reform can look very different in different places, and different reforms can end up being implemented in similar fashions. Here there would be a place for qualitative or observational techniques. (It is interesting that nearly everything mentioned in the two papers is either an input or an outcome variable, but there is little mention of the processes linking the inputs, including policies, to the outcomes. That mirrors the strengths of NCES—much progress in developing information and indicators on inputs and outcomes, but relatively weak on measures of process, i.e., what goes on in the black box, in the classroom.)

What Is the Effect of Reform?

This is the toughest and most problematic type of information for NCES to gather. The agency is not in the business of program evaluation. There is an important difference between monitoring what is happening in education and evaluating those happenings. The first is an appropriate role for a federal statistical agency, while the second is not. Furthermore, establishing the impact of a particular program is difficult and complicated, and large-scale national surveys such as those that NCES typically conducts are not well suited to doing such evaluations. The difficulties are illustrated by an NCES publication released 3 years ago.

One provision of the Hawkins-Stafford Amendments of 1988 reauthorizing NCES was a mandate to study the effects of higher standards (as the result of reform) on student enrollment and persistence, academic achievement and graduation rates. In the end, the report consisted primarily of two types of information: an enumeration of the types of reforms raising student standards that had been enacted between 1984 and 1990 and the number of states involved (the first category of information, how much reform is happening); and secondly, a description of trends in student outcomes over the same period. The report made it very clear, however, that one could not link the two types of information together in a causal fashion. The two sets of events had occurred during the same time period, but one could not conclude that one caused the other. The Executive Summary of the report emphasized that point strongly in the following passages (Medrich et al. 1992, pp. vi and vii):

> Even though the states are increasingly active in defining student standards, linkages between these initiatives and student outcomes are difficult to measure for a number of reasons:
>
> - States have adopted different reforms at different times, and no two states have adopted the same exact requirements;
> - Even in cases where similar types of reforms can be identified among several states, there is much variation in how these initiatives have been implemented from state to state;
> - While some reform activity occurs at the state level, far more occurs at the school district, school, and classroom levels; and
> - Over time, demographic shifts have been dramatic in many states, and it is difficult to control for the effects of reform, over time, on different populations.
>
> Although it may be possible to ascertain whether changes in student outcomes have occurred in a positive direction over time, this only suggests that state reforms may be associated with these outcomes. Given the caveats noted above, linkage in a statistical sense cannot be substantiated . . . .
>
> . . . [I]n order to establish linkages between state reforms and student outcomes, it will be necessary to examine in more detail the ways in which states implement reforms (the translation from policy to practice) and the extent to which reforms change practice; the impact of specific reforms on local school districts and classrooms; and changes in curriculum content and the quality of instruction associated with, or resulting from, reforms of student standards.

Emerson Elliott has emphasized that there is an important distinction between research and statistics. This is an area far better suited to research—where one can gather pre-reform data, study implementation, gather information about all the contextual factors involved, and look at the "value added" by reform mentioned in the Cross and Stempel paper—than to large-scale data collection.

What Kinds of Contextual or Baseline Data Are Needed?

The last set of topics involves areas where the authors felt data that could serve as important contextual or baseline information for policymakers interested in reform were not available. NCES already collects some of these types of information, but may need to collect more or make the availability of such information more widely known. For example, the longitudinal studies and NAEP are mentioned in the papers, but not SASS. NAEP was suggested as a way to gather more information on teachers' education and development, but SASS already collects a great deal of such information. The agency may need to do a better job of making people aware of SASS and the type of data it collects. In other cases, NCES is exploring ways of gathering the kinds of data mentioned, most notably, postsecondary assessment data. For some other topics, NCES has considered the need for such data, but is not actively pursuing ways to collect them for a variety of reasons, including cost.

Strategies Related To Data About Reform

The other broad theme in these papers is strategies that NCES might pursue in a variety of areas, including data collection, what to do with the data once they have been collected, and dissemination. The following discusses a few of these that I found interesting.

Jennings and Stark suggested preparation of an annual report on reform, an "encyclopedia of reform," that would include how much of various kinds of reform was going on, with information by state. The report would include not just NCES-generated information, but also data collected by others. They suggest the information could be put up on the Internet and updated as new information became available, rather than waiting to release all of it simultaneously in the publication. Such a report is an interesting suggestion, but probably would be more appropriate for another organization, such as OERI's Office for Reform Assistance and Dissemination (ORAD) or possibly the Goals Panel (if it survives), which is already doing a considerable amount of this type of activity.

In terms of NCES using data that have been collected by others, which is discussed at length in Fritz Scheuren's paper on administrative record data, I have one concern. NCES needs to be very careful about the caliber of such data that it releases, either electronically or in publications. Utilizing and releasing such data, despite being issued with many caveats, will be seen by the outside world as an endorsement of them by NCES. How this is done, including whether NCES imposes some standards the external data must meet, will affect the likelihood that such a strategy could compromise the reputation of NCES for accuracy, reliability, and impartiality.

The papers also emphasize the need for more analysis of the data NCES collects and production of reports that are more attuned to the audiences of policymakers and the public. NCES is already vigorously pursuing these strategies. The new Education Statistics Services Institute (ESSI) should facilitate doing more and doing it better in these areas and the agency would welcome additional concrete suggestions. The trick is to produce material that is policy-relevant, but does not cross the line into policy evaluation or policy recommendations.

It is also intriguing to think what a paper written by a researcher interested in reform would have looked like. These two papers reflect what policymakers are interested in right now. What is discussed is a lot of separate topics, representing breadth, but not necessarily depth, of information. My guess would be that a researcher addressing this topic would have focused on in-depth analysis of a few topics or on a structure to monitor change, apart from the current "hot topics" and would have come up with suggestions for systematic, in-depth studies, either as new surveys or as components or modifications to existing surveys.[2] For example, they might have proposed a longitudinal study of schools as a component of SASS. (Such a survey might be good for gathering three of the four types of information identified in these papers, but not for the effects of reform. It could be used to track what kinds of practices and policies were pursued in a group of schools over time, but to collect information on the effect of those practices would require collecting information on students, with before and after reform measurements, as well as much contextual information.)

The difference between that sort of suggestion and what is in these papers reflects the differing demands of the various audiences of NCES. While everybody wants more information, researchers tend to want in-depth, systematic studies, while policymakers are looking more for breadth rather than depth for issue-brief type of information that is readily accessible, i.e., for sound bites. Part of the challenge for NCES is to provide useful information to both types of audiences in a time of fiscal constraint.

One strategy I feel needs attention, but which was mentioned in neither paper, is how NCES can do a better job of identifying (and gathering information about) new issues that are emerging on the horizon. It takes a very long time to implement new items on existing surveys (and even longer to mount new surveys or survey components). Are there ways NCES can identify new issues earlier and collect information on them before building them into large-scale surveys? The FRSS is one option to collect data of this sort, but there may be others. What might serve as an "early warning system" for identifying potential upcoming information needs?

Summary

In summary, I see these papers as stimulating NCES to take a broader view of its role. NCES will not pursue all of the topics mentioned in the papers; that is not fiscally possible nor appropriate. However, NCES could think about who might provide such information to the American public (and how this might occur), and could play a role in seeing that it happens. The NCES role would vary by topic and activity, sometimes serving as a facilitator or coordinator. That would involve a new role for NCES, acting as a broker of information, identifying holes, and getting others to fill them rather than doing it directly (nor would NCES necessarily have to release the information, but rather monitor that someone does). It also implies working more closely with a wider group of actors, starting with colleagues in other parts of OERI and the Department of Education, but also reaching out to private groups, associations, foundations, business, and other interested parties.

Notes

1. Another is home schooling. In the October 1994 Current Population Survey (CPS), items were included on home schooling. Out of the nearly 60,000 households in CPS, there were about 100 children between the ages of 6–17 who were being home schooled.

2. A paper submitted to NCES subsequent to the Futures conference (Baker 1996) looks very much like this. In making recommendations about changes to SASS that would help monitor reform efforts and their impacts, the author suggests reorienting SASS so that its primary focus would be on gathering information on schools as organizations. He cautions against focusing on

particular "reforms," because they come and go (and come again) relatively quickly. "[T]he key is to think of ways to capture information about reform without being tied to any one particular trend over a lengthy time" (p. 31).

References

Baker, D. 1996. "Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with Comments on School Reform, Governance, and Finance." Commissioned paper for the U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Forum on the Design of the 1998–99 Schools and Staffing Survey (SASS).

Medrich, E.A., Brown, C.L., Henke, R.R., Ross, L., and McArthur, E. 1992. *Overview and Inventory of State Requirements for School Coursework and Attendance*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

APPENDIX A

Topics and Strategies Mentioned in Two Papers

Jennings and Stark	Cross and Stempel
Topics	
Effect of reform on achievement	Effect of reform on achievement
	Value added by reform
Standards-based reform—what states are doing (updated), standards and assessments, OTL	Effect of academic standards on schools
Graduation requirements	
Charter schools—how successful	Charter schools—evaluate performance
Reform networks	
School choice	
Private, for-profit companies run schools	
State takeovers	
School finance, state-by-state	
Effect of reduction in federal funding	
Effect of going to block grants	
School facilities	
Vocational education—tech-prep, school-to-work	
Home schooling—state-by-state, oversight procedures	Home schooling
Postsecondary reform initiatives	
Postsecondary student achievement	
Outcomes other than achievement	
International comparisons with states	
Regional, state, local data on reform	
Information on demographic, economic characteristics of school districts	
Report on specific reforms at all levels	Degree to which reforms are being implemented
	International benchmarking
	Motives and expectations of other countries
	Coordination of curriculum and expectations within schools and across schools in district
	Student mobility
	Interaction of elements of reform
	Technology—use, infrastructure
	Magnet schools
	Vouchers

Jennings and Stark	Cross and Stempel
	Site-based management—many different kinds; impact on learning
	What occurs in classroom
	Quality/uniformity of assessments/testing in classroom; data on teacher grading schemes
	Cost/benefits of various reforms
	How successful schools streamline use of limited resources
Strategies	
Broadest definition of reform to remain impartial	
Advisory committee—what reforms; priorities	
Annual report—"encyclopedia of education reform," state activities, include data of others	
Other sources of financing—i.e., business and foundations	
Compile information collected by others; "repository" of others' data	
Other data on achievement—state assessments, college entrance exams	
Use state data for sub-state data	
Collect and report by state	
Put information up on Internet—includes research findings and updating data	
Make data widely available to researchers and others	
Determine what is education reform; common definitions	
Timely information, relevant, and easily accessible	
Concise reports suitable for policymakers	
Have information available for reauthorization; time publications to that schedule	
Most recent data on Internet	
More analysis of data—more interested in reasons than mere facts; why reforms work; need data over time	More analysis of data
Encourage analysis of reform data by researchers	

Jennings and Stark	Cross and Stempel
NCES/OERI conference on how to make data more accessible and promote analysis	
Contract out analysis	Integrate qualitative and quantitative information about reform
	Modify teacher part of NAEP to get information about teacher education and development (SASS)
	Longitudinal information on teacher careers
	Collect data from multiple sources; unreliability of self-reports
	Independent data collection agents or "inspectors" because self-reports are unreliable
	Link information at all levels to see what is happening in specific classrooms
	Extend purpose of data, more reliable, more "subtly realized"
	NAEP—expanded items to examine schools' role in achievement
	Data collections more delicate, sensitive, focused

3

Curriculum, Pedagogy, and Professional Development

Enhancing Opportunity to Learn Measures in NCES Data

Dominic J. Brewer
Cathleen Stasz

INTRODUCTION

What takes place in American K–12 classrooms? What is being taught, how is it being taught, by whom, and with what resources? Knowing the answers to these questions would seem to be a necessary information base upon which to build public policy aimed at boosting student performance, ensuring an equitable delivery of schooling for all students, and guaranteeing accountability of teachers and schools. The nation's school system has been the focus of much public discontent over the past decade, centered on perceived declines in student academic achievement, school inefficiency, and lack of accountability. Consequently, schools have been subjected to an unprecedented era of "reform," ranging from changes in assessment to new curricula and graduation requirements to new models of school organization. Most of these changes are ultimately designed to bring about improvements in student outcomes, typically measured (narrowly) by standardized tests, via changes in what takes place within classrooms around the nation. However, given limited understanding of the determinants of student performance, the difficulties of measuring the inputs, processes, and outputs of schooling, and the many and disparate activities and clientele of the school system, systematically assessing the real impact of these reforms is no easy task. A precondition for this, however, is an accurate, detailed picture of what takes place in American classrooms.

A vast volume of research within education and other disciplines has attempted to map out and explain the processes of teaching and student learning. Within the classroom setting, the focus has been on curriculum content, pedagogical strategies and instructional goals, teacher characteristics, and other instructional resources. Recently, national data on such issues, based primarily on survey responses of teachers, have been collected by the National Center for Education Statistics (NCES). Typically part of large-scale national (often longitudinal) studies designed to meet a variety of diverse needs (e.g., researchers from many disciplinary backgrounds and policymakers interested in a host of issues), these data have a broad rather than a deep focus. They do, however, have the advantage of drawing on large sample sizes and are carefully designed and implemented. While this effort to collect data on curriculum and pedagogy at the national level is in its infancy, there is considerable doubt as to whether the complex nature of teacher and student behaviors, and their interaction in a classroom setting, can be captured by survey data.

This paper reviews attempts to date to collect classroom-level data, and discusses whether the mapping of the intricate, multidimensional activities of the classroom can be improved via better survey designs and instruments, or via other forms of data. The next section defines in more detail what we mean by curriculum and pedagogy, utilizing the concept of "opportunity to learn" (OTL). The following section discusses the rationale and uses for such data, and briefly outlines possible future needs in this area. The paper then presents an overview of existing NCES data collection efforts via national surveys. Several non-NCES major data collection efforts have been undertaken in recent years geared toward improving measures of OTL through a variety of alternative methods. These include the use of teacher daily logs, collection of classroom written assignments, test and texts, teacher interviews, classroom observation, and videotaping of classes. While this work is relatively new, it provides some potential avenues for future NCES data gathering. The paper concludes with a set of recommendations with regard to future efforts.

DEFINING "CURRICULUM AND PEDAGOGY"

We take an expansive view of "curriculum and pedagogy," focusing on what takes place inside classrooms at the K–12 level in the broadest sense. This includes what is being taught (e.g., curriculum content); how it is being taught (i.e., pedagogical strategies); who is teaching (e.g., teacher—and sometimes student—characteristics); the instructional resources being used directly (e.g., textbooks); and the resources for teachers that support instructional goals (e.g., planning time, staff development, and opportunities for faculty collaboration). The activities of any classroom thus include not only the behaviors of teachers but also the activities and interactions among students, and between teacher and students. This whole gamut of classroom intentions, behaviors, and activities is obviously an extremely complex one with many dimensions. Therefore, obtaining a coherent and usable description of this picture is no easy task.

Opportunity to Learn (OTL)

One widely used way to organize thinking about curriculum and pedagogy is the concept of "opportunity to learn" (OTL).[1] Typically OTL research has divided classroom attributes into three distinct categories: curriculum content, instructional strategies, and instructional resources. Briefly, curriculum[2] refers to "the knowledge and skills the teacher presents for the students to learn during their classroom experiences" (NCES 1995b). This typically includes major and minor topic coverage, time spent, and more subtle teacher emphasis on topics. Instructional strategies refer to the ways in which teachers convey material to and engage their students. Traditionally, the latter has included issues such as the manner in which material is presented to students (methods, pace); questioning strategies; communication with students; expectations for students; classroom organization; grading and homework policies; allocation of time within a class period; and content organization. Instructional resources include basic learning materials such as books and supplies; equipment (e.g., computers); and the physical classroom environment (such as heat, light, furniture, and so on). We also include teachers' knowledge and preparation of curriculum and strategies under this heading (sometimes it is classified under curriculum). For more complete descriptions of the elements of OTL, see NCES (1995b).

While the OTL framework is a useful one, we believe it is important not to view the three aspects of OTL as rigid. Indeed, developing a richer and fuller picture of classroom activities hinges on being able to successfully identify and measure the *interactions* and overlap between curriculum, pedagogy, and resources, and their effects on learning. It should be clear that many classroom activities defy simple categorization; for example, as Porter (1991) has pointed out, separating curriculum content from pedagogy is tricky: "If a teacher uses story problems to teach problem solving, but all of the story problems involve the same format and the same operation for solution, then after the first few problems, the task becomes one of drill-and-practice for skill, not one of application and problem solving" (p. 18). Similarly, Burstein et al. (1995) note the interaction between content and teaching strategy and inability of survey data to capture "subtle differences in how teachers define and used different techniques" (p. 36). Thus, while one teacher might draw heavily from the textbook and do most of the talking in class, and another might use other sources and engage students in lively exchanges, both are likely to report these activities as lecturing.

Student Learning, Teacher Doing

In setting our focus on classroom activities in broad terms, we deliberately wish to call attention to what we believe is a neglected part of many analyses of curriculum and pedagogy: *the process of student learning*. NCES and most other data collection activities emphasize assessment of *teacher* behavior whether that stems from traditional process-product research or from newer, "reform-oriented" pedagogical approaches, such as the NCTM standards (Burstein et al. 1995). Clearly, however, teachers represent only one part of the classroom. This emphasis on teaching is consistent with instructional design theory (cf. Reigelut 1987). From this perspective, the task is to design an instructional delivery system that transmits content and skills in a clear, well-structured, and efficient manner. The approach stems from behaviorist theories of learning, but also has assimilated aspects of cognitive research in recent years (Collins in press).

In contrast to an instructional delivery view, a constructivist view argues that education should help students construct their own understandings. This perspective leads to an emphasis on learning rather than teaching, and on facilitative environments rather than instructional goals. It implies an approach to education that looks very different from traditional instructional design theory, and where considerations of curriculum content, for example, do not hold center stage. Collins' discussion (in press) of design trade-offs among learning goals, for example, addresses goals for what students should learn: memorization versus thoughtfulness, whole task versus component skills, breadth versus depth of knowledge, and so on. Such considerations about student learning goals apply whether the topic is math, social studies, or electronics.

Recent NCES surveys (e.g., The National Educational Longitudinal Study of 1988, [NELS:88]), as well as non-NCES efforts (Burstein et al. 1995), include teacher survey items based on reform efforts that in turn reflect constructivist views of education (e.g., National Council of Teachers of Mathematics [NCTM] standards, and California curriculum frameworks). While these items (discussed in more detail later) are meant to assess teachers' use of newer practices, they still focus on what *teachers* do.

While recent OTL-oriented studies have recognized and attempted to rectify the failure to assess student learning (Smithson et al. 1995), these efforts are still too new to determine if they validly represent student learning processes. Studies of how students experience the curriculum or how students think about what teachers do seem to be one starting point for examining the student learning process. Unfortunately, we currently know little about the varieties of student experience in classrooms, and what we do know has not been integrated into recent theoretical orientations toward teaching and learning (Good 1995; Erickson and Shultz 1992).

WHY COLLECT NATIONAL DATA ON OPPORTUNITY TO LEARN?

Why collect national data on opportunity to learn? We believe obtaining basic information about the nation's classrooms is a useful, necessary, and important undertaking for both policymakers and the research community. The extent to which we are able to describe the current educational system is surely a crucial element in efforts to improve it (Stecher 1992). Several reasons for this position are discussed in this section of the paper. For example, fundamental questions about effective practice—what teacher behaviors promote student learning—remain unanswered. Policymakers interested in reaching informed decisions about school reform need to know how opportunity to learn is distributed across students and with what effect, and how much change at the classroom level really takes place as a result of ongoing reforms in curriculum, assessment, and teaching practice. Further moves toward increased teacher and school accountability may require more numerous and more refined indicators of classroom activity. Without a clear view of *why* data on curriculum and pedagogy are needed, and how these needs may change, future data collection is unlikely to generate useful insights and may divert needed resources from elsewhere. As Porter notes, "the value of an indicator of school processes is determined by its problem orientation and policy relevance" (1991, p. 23).

What Promotes Student Learning?[3]

Prompted by unfavorable international comparisons, the overarching interest of policymakers over the past decade has been to improve American students' educational attainment, typically as measured by standardized test scores. Following a lengthy period in educational research and practice in which "inputs" to the schooling process and how these inputs were distributed across types of students were the focus of attention, emphasis switched in the 1980s to a strongly outcome-oriented paradigm. Raising educational productivity is clearly an important goal for the nation's schools. To the extent that researchers could identify the ingredients of successful classrooms—attributes of teachers or curriculum or classroom resources, for example—which are generalizable across different settings, a formula for improving student outcomes would be found. This "recipe" could then be applied throughout the nation in order to improve outcomes.

In fact, this line of research, either in the "process/product"[4] or "effective schools"[5] genres within educational research, or in the "educational productivity"[6] literature within economics, has been unsuccessful in arriving at strong conclusions in regard to "what works." Process-product research has revealed, for example, correlations between the pace of instruction, how information is presented and teacher's questioning strategies, and student outcomes. Work on curriculum content has generally shown some link between the number of courses taken in a particular subject and achievement in that subject. The effective schools' literature suggested that "instructional leadership," clear school goals, and high expectations were particularly important in promoting student outcomes. The productivity paradigm has led to fewer robust conclusions, though there is a dispute over the interpretation of the numerous studies. In general while crude measures of teacher "ability" have been shown to be important determinants of student achievement, holding a set of school and student characteristics constant, other indicators such as teacher experience and degree levels have shown an inconsistent relationship to test scores.

The reasons for the rather disappointing results from this research are numerous. First, "outcomes" are much broader than standardized tests, which are often poorly designed and inadequate measures of student learning. Since outcomes such as problem-solving skills, self-esteem, communication skills, and citizenship are far harder to assess in a systematic and measurable fashion, they have tended to be ignored or relegated to footnotes by researchers. Pressure from the public to achieve tangible gains in measured performance has led to test scores being used as "political symbols as much as assessment devices" (Good 1995, p. 4).

Second, much of the process-product research has been based on small unrepresentative samples and has focused on one particular factor or part of classroom activity; in general, it has utilized univariate statistical methods in which causality is impossible to infer with confidence. Available national data for use in more sophisticated statistical studies, while improving, remain crude. Only recently, for example, have student tests on national surveys been explicitly tied to curriculum content and their teachers and classrooms. Most of this research has been conducted by economists who have focused almost exclusively on resource inputs (such as classroom size, or expenditures per student), and has ignored the more subtle and intangible multitude of teacher behaviors (such as whether a particular teacher lectures, or teaches a certain number of units of algebra). There have been few attempts to combine the two strains of research to produce a firmer basis upon which to draw conclusions as to "what works" in terms of classroom activities (see Murnane and Phillips 1981).[7]

Third, and perhaps most fundamentally, it is a debatable point as to whether such outcome-oriented analyses can provide a recipe that educational policymakers could adopt given differences across schools and classrooms. While it is possible, for example, to identify student and teacher classroom characteristics and crude indicators of curriculum and instructional strategies, the ways in which these elements interact in any given classroom on any given day are likely to be very complex. Indeed, some researchers have argued that the multidimensional nature of both inputs and outputs in the classroom makes this line of research impotent. If idiosyncrasies dominate, a formula that could be implemented across classrooms and schools will never be found (Monk 1992).

While these problems associated with answering the question "what promotes student learning?" have led to few strong conclusions, it is probably premature and overly pessimistic to abandon this line of research altogether. Indeed, more recent work focused on the impact of curriculum content on achievement has been more successful (see the brief overview in Burstein et al. 1995, pp. 3–5). Increased attention on data needs in the curriculum and pedagogy area is partly a result of earlier research failures and a desire to get inside the "black box" of schooling in a systematic way with nationally representative data. There seems to us to be further opportunities to combine modern statistical techniques with improved and more refined national data on classroom activities and resources. Although it seems unlikely that a "silver bullet" will be discovered that can be used to cure the nation's educational productivity problem, even if better measures were to be collected by NCES, there is some merit in attempting to further refine measures of classroom activities in order to gain a better understanding of the ways in which curriculum content and instructional strategies may be related to student outcomes.

Do All Students Have an Opportunity to Learn?

Although somewhat de-emphasized in recent policy debates, descriptions of what takes place in the nation's classrooms are useful from an equity perspective. Ensuring educational equity is an important goal for the nation's schools. One particular emphasis in recent years has been on developing "indicator systems," a key component of which is information on school processes (see Shavelson et al. 1987; Stecher 1992; McDonnell 1995). The main purpose of these efforts is primarily to provide data to policymakers (Porter 1991). OTL provides a way to determine if different students have equal educational opportunities. Equal opportunity is a fundamental concept in public schooling in the United States. To the extent that a certain allocation of educational opportunities across students is desirable, and that educators should strive to achieve such an alignment, it is necessary to first provide a description of the distribution of OTL. The growing interest over the past decade in obtaining indicators of classroom processes stems partly from a recognition that crude input measures are inadequate for assessing educational opportunity. Traditionally, resource (dollars per pupil) differences across school districts (Kozol 1991) have been emphasized, given the perceived link between spending and student outcomes. As is well known, this has led to numerous attempts in many states to equalize spending where inequities exist led by the California *Serrano* court decision in 1971 (*Serrano v. Priest* 1971). Equalization, such as it has occurred, has not led to an equalization of educational opportunity.

OTL offers a much richer process-oriented description of the schooling students receive. For example, the concept is based, in part, on a link between students' curricula exposure and their achievement (McDonnell 1995). Clearly, if students are not taught particular mathematics topics, for example, they cannot hope to score well on tests of those topics. NCES recently examined the link between course taking and achievement in math and science for students from different social backgrounds (Hoffer et al. 1995). Hoffer's analysis found that students from higher socioeconomic status (SES) families complete more courses in these subjects. In addition, students who complete more math and science courses show greater achievement score gains during high school, regardless of gender, race–ethnicity, and SES. Thus, additional coursework pays off for all students (see also Jones et al. 1986). The demonstration of a link between curricula exposure and outcomes has not been lost on policymakers; for example, Goals 2000 gives inducements to states to establish curriculum and student performance standards (Burstein et al. 1995, p. 5). Mapping course-taking patterns, then, is an important way to assess opportunity to learn. Further, there is already a good deal of evidence demonstrating considerable inequity in the distribution of classroom opportunities. For example, low SES and minority children tend to be taught by less qualified teachers, and have less access to tangible instructional resources such as computers (see Guitton and Oakes 1995, p. 324, for citations).

Distribution of opportunity to learn remains an important issue for public policy. It is not clear whether we know, for example, if "opportunities to learn are significantly different for students in Seattle as compared to those in Indianapolis" (McDonnell 1995, p. 310). On a more subtle level, information about the distribution of curricula offerings or instructional practices has led to efforts to reorganize course taking and classroom organization, for example, in "de-tracking" reform efforts. Several studies have shown that curriculum decision making in schools can match students with curriculum in ways that limit course taking and have implications for what students are exposed to and learn, and in some cases, their future educational opportunities (Oakes et al. 1992). In many states, increasing immigration raises concerns about the school's ability to cope with the educational needs of immigrant students and to keep them in school. OTL-type measures of what takes place within classrooms are likely to be informative about developing educational policies that can most effectively help such children. While use of OTL indicators for assessing equity is not without problems (as Guitton and Oakes [1995] argue, for example, different conceptions of equity lead to rather different emphases in developing OTL measures), it does provide an additional rationale for collecting data on curriculum, pedagogy, and instructional resources.

Are Current School Reforms Being Implemented?

A goal of many ongoing reforms being implemented in America's schools is to change curriculum and pedagogy in the nation's classrooms. It is difficult, however, to know whether reform rhetoric is translated into measurable change at the micro level: how change is implemented, the speed of implementation, and the barriers to change. Implementation of these reforms is a highly localized endeavor. Therefore, one important purpose of national data collection efforts on curriculum and pedagogy that is particularly important in an era of unprecedented change in schools is as a mechanism for monitoring that change at the level where it is being implemented.

In principle, descriptive information can provide a means of mapping out the type of changes that are taking place in schools, and where these are occurring. Further, with sufficient detail and coverage, survey data can be used to determine if indeed reforms are translated into positive discernable outcomes via non-experimental analyses. The National Assessment of Vocational Education (NAVE) provides one example where survey data have been used to track national reforms in vocational education. The most recent NAVE examines the extent to which states and localities are implementing specific program improvements mandated by Congress in the 1990 Amendments to the Carl Perkins Act (NAVE 1994).[8]

However, tracking reform via national surveys is no easy task given that change takes time and implementation is often slow. Surveys can only provide a snapshot of what is taking place at a point in time. Hence, it may be difficult to monitor change during a transition period; simply finding the appropriate words for survey items that adequately describe what should be taking place is problematic (Burstein et al. 1995). Given the pressure for certain reform it is impossible to know without additional information whether respondents simply adopt the current popular view about what they *should* be doing. This may be the case, for example, with the current national standards movement. Some of these issues are considered elsewhere.

Inherently tied to current reforms and the outcome-based emphasis of the past decade are increasing calls for accountability of teachers and their schools. Blank (1993) stresses that the focus of education reformers on accountability has increased interest in indicators of school processes (not the least via committees sponsored by the National Science Foundation, National Research Council, and NCES, among others). Hence, one additional rationale for collecting data on a national level on OTL is what McDonnell (1995) calls "high stakes" uses. OTL standards might be used, for example, in assessing whether or not schools meet certain practice standards and hence may be subject to sanction, i.e., using "force action in prescriptive ways" (Guitton and Oakes, 1995, p. 325). This view has also been expressed by O'Day and Smith (1993). At the very least they might be used in conjunction with other data in any system of school rewards based on outcomes (student performance). Clearly, any measures used for such purposes require a high degree of reliability and widespread acceptance from educators as legitimate measures. Hence, the use of OTL-type measures for accountability purposes has generated considerable controversy (Porter 1995; McDonnell 1995; and Guitton and Oakes 1995). NCES data are unlikely to be used for such purposes.

EXISTING NCES DATA: SCOPE, METHODS, AND LIMITATIONS

NCES and its predecessors have collected limited data on curriculum and pedagogical practices in American schools for some time.[9] The primary mode of data collection has been to include items relating to classroom activities on large national surveys completed by principals or other school administrators and by teachers. These items have been primarily limited to information about teachers' educational backgrounds, and school or classroom resource indicators (such as number of pupils per teacher). Some recent major NCES surveys include High School and Beyond (HSB), The National Educational Longitudinal Study of 1988 (NELS: 88), the Schools and Staffing Survey (SASS), and the ongoing National Assessment of Educational Progress (NAEP).[10] Detailed information on classroom processes including curriculum content and instructional strategies is relatively new to NCES surveys; NELS was the first major survey to contain a considerable set of questions relating to specific curriculum topics within major subject areas and items focused on teaching behaviors within a particular classroom. Stecher, in reviewing available measures of curriculum content in 1992, concluded that very little was known about topic coverage or instructional methods (1992, p. 56). Further, only NELS has systematically combined longitudinal classroom-level survey data on students and their teachers that include student test data.[11] Hence, national data collection in this area must be considered to be in its infancy.

Typical NCES teacher and school administrator survey coverage is summarized in Table 1. This table is not intended to be exhaustive but provides an illustrative introduction to data on OTL. We have categorized items into three groups based on an OTL framework: curriculum content, instructional practice, and instructional resources. The four major surveys noted above are included in the table.[12] In general, the table illustrates that only NAEP and NELS contains classroom-specific data on detailed curriculum topic coverage and teachers' instructional strategies. Even instructional resource measures have been limited and confined to the school rather than classroom level, obscuring variation between classrooms within a school.

Table 1—Items on opportunity to learn in selected NCES national teacher/school surveys

	Curriculum content	Instructional practice	Instructional resources
I. National Assessment of Educational Progress (NAEP) Annual 1969-1980; then biennial; student and teacher components	Broad topic emphasis within subject area planned for academic year for each class; type of skills taught	Broad instructional methods, type and frequency of student tasks assigned during class; amount of homework; type and frequency of assessment; use of resources	Class size; access to resources (e.g., calculators); teacher education and training, including in subject-specific content areas and teaching techniques
II. High School and Beyond (HS&B) Longitudinal student surveys 1980 (10th and 12th graders), 1982, 1984, 1986, and 1992 (sophomores only); school survey (1980, 1982); administrator/teacher survey 1984	None	General teaching goals, class time allocation, homework and assessment strategies, student recognition	Average class size and ability, school level expenditures/pupil, teacher credentials
III. Schools and Staffing Survey (SASS) Teacher, administrator, and school surveys 1987–88, 1990–91, 1993–94	Broad topic areas within subject for each class taught	None, general attitudes/perceptions toward teaching	Class size, ability level for each class taught, otherwise none
IV. National Educational Longitudinal Study of 1988 (NELS: 88) Longitudinal student surveys 1988 (8th grade), 1990, 1992; teacher surveys (tied to students) 1988, 1990, 1992; school surveys 1988, 1990, 1992; parent surveys 1988, 1992	Topic emphasis within subject area; general concepts specific to subject area	Type and frequency of instructional method used; allocation of time to whole group	Class size; teacher credentials, including coursework in subject areas

SOURCE: Compiled by authors from respective survey instruments. See text for discussion.

HS&B was a national longitudinal study designed as a followup to the National Longitudinal Study of the High School Class of 1972 (NLS–72). The basic HS&B data consist of two cohorts (sophomores and seniors) of high school students initially surveyed in 1980. Basic school information was provided by principals, including some information on general school policies and instructional resources (class size, dollars per pupil, teacher credentials, and so on). A supplementary Administrator and Teacher Survey was conducted in 1984 that included separate surveys for principals, guidance counselors, and teachers. While teachers were not linked to particular students or asked about a particular class, they were asked a number of questions regarding the extent to which they controlled their classroom resources, curriculum content, and teaching techniques; how often classes were interrupted "on an average day"; the importance of general goals (such as "academic excellence," "good work habits," and "discipline," and "moral or religious values") in their teaching; allocation of classroom time between "daily routines," "getting students to behave," and "instruction or student practice of skills"; and detailed items relating to their qualifications and background. Clearly these items are unspecific and general in nature.

SASS consists of a set of surveys conducted in 1987–88, 1990–91, and 1993–94. The latest wave consisted of a teacher questionnaire sent to 65,000 teachers, along with administrator and school components. No student data were collected. The teachers were asked about classes taught during the most recent full week at the time the survey was completed. Broad curriculum topics—for example within mathematics 11 topic area codes such as "general mathematics," "trigonometry," "calculus"—class size, grade level, and ability level were included, but no specific items on instructional strategy. Teachers were asked about their backgrounds for teaching specific subjects.

NAEP's "The Nation's Report Card" was started in 1969 as a way of tracking the educational performance of the nation's school children. Students in different grades are tested in a variety of subjects and their background information is collected. In addition, their teachers are surveyed to gather data on their background and instructional practices. NAEP continues to be conducted every 2 years, with samples of 4th, 8th, and 12th graders being tested since 1988; math and reading are assessed every 2 years, science and writing every 4 years, and other subjects less frequently (NCES 1995a). Curriculum content is surveyed with broad indicators for each subject. For example, 8th-grade mathematics teachers were asked in the 1992 NAEP: what emphasis they *plan* to give ("heavy," "moderate," or "little") to five topic areas during the course of the academic year: "numbers and operations," "measurement," "geometry," "data analysis, statistics, and probability," and "algebra and functions." Instructional strategies are also assessed via one or two items. For example, 8th-grade writing teachers were asked in the 1992 NAEP: "Do you use any of the following instructional approaches?" "Grammar or skill-based instruction," "writing process instruction," "integrated reading and writing," "writing about literature," "writing across other subject areas." A threefold scale ("Yes, as a central part of instruction," "Yes, as a supplement to instruction," and "No") was utilized.

NAEP also assesses student activities. On the 1992 NAEP, for example, 8th-grade math teachers were asked: "How often do the students in this class do each of the following things?" Eleven activities were assessed—such as "do a mathematics problem from their textbooks," "do a mathematics problem on worksheets," "solve mathematics problems in small groups," "discuss solutions to mathematics problems with other students"—on a scale of "almost every day," "once or twice a week," "once or twice a month," or "never or hardly ever." Interestingly, questions relating to these issues are not confined to teachers; for example, *students* are asked: "How often does the teacher . . . read aloud . . . do a problem on the board?" In the past, however, NAEP has sampled students and teachers in such a way that only a subsample of a teacher's students were sampled.

NELS is perhaps the most detailed and most recent collection effort. A group of 1988 8th graders have been surveyed and tested in 1988, 1990, and 1992, along with detailed survey information from these students and their teachers in at least one subject-specific class. Additional school-level and parent information was also collected. NELS is thus unique in that it permits researchers for the first time to link a national sample of students with the teachers who actually taught them. Further, the standardized tests administered to the students were, in contrast to HS&B, for example, linked to the curriculum studied by the students. The teacher surveys contain detailed information on the curriculum content used in a particular class, instructional methods used, and the teacher's goals for the class, in addition to the standard range of items on teacher educational background, class size, and school-level resource measures. Some examples of specific items taken from the 1992 mathematics teacher survey are given in Table 2 in order to provide an indication of the level of detail of these survey items.

The curriculum content items include both specific topic coverage and degree of emphasis on different objectives for mathematical learning. The list reflects newer conceptions of appropriate instructional goals for teaching mathematics (e.g., learning to represent problem structures in multiple ways). Similarly, the instructional practice items attempt to discern the percentage of time engaged in teaching (individuals, groups, whole class, labs) relative to other non-instructional activities, such as maintaining order. Items assessing teaching methods attempt to distinguish between teacher-centered activities (e.g., lecture, lead group discussion), and more student-centered activities (e.g., give oral reports, work in cooperative groups), and the frequency of these activities and of uses of instructional media. Like the content and emphasis questions, these methods questions appear to reflect "reform" practices espoused by mathematics educators in such documents as the NCTM Professional Standards (1991).

National survey data collected by NCES have several major benefits: sample sizes are typically large, nationally representative, and carefully designed and implemented. Hence, these data have a degree of generalizability that other research efforts cannot match. The data collected by NCES contribute in large part to the goals of data collection in this area outlined earlier. For example, NELS affords researchers the opportunity to explore the relationships between curriculum content, instructional strategies and resources, and student achievement in a particular subject area, while controlling for school and other contextual factors and students' prior ability level. Hence it is a remarkably rich data source. NAEP provides a reasonably comprehensive snapshot of the curriculum in math and reading across the nation. However, NCES data collection efforts are geared toward the many diverse groups that the agency is charged with serving, resulting in a broad-brush approach, rather than one focused more deeply on specific areas such as curriculum content at the K–12 classroom level.[13] Hence curriculum content indicators, while becoming more detailed, remain at a relatively general level with a handful of topic areas identified within a subject. Instructional behavior is captured in crude terms both in terms of strategies identified and the response scales used. Surveys have major limitations for collecting very rich information on curriculum and pedagogy.

Table 2—Examples of teacher survey items from NELS Second Follow-up

Curriculum Content

Have you taught or reviewed the following topics in this math class during this year?
Integers; patterns and functions; linear equations; polynomials; properties of generic figures; coordinate geometry proofs; trigonometry; statistics; probability; calculus.

Scale: No, but it was taught previously; Yes, but I reviewed it only; Yes, I taught it as new content; No, but I will teach or review it late this school year; No, topic is beyond the scope of this course.

How much emphasis do you give to each of the following objectives?
Understanding the nature of proofs; memorizing facts, rules, and steps; learning to represent problem structures in multiple ways; integrating different branches of mathematics; conceiving and analyzing the effectiveness of multiple approaches to problem solving.

Scale: None, Minor, Moderate, Major

Instructional Practice

What percent of your time did you spend?
Instructing whole class; instructing small group; instructing individuals; maintaining order; administering tests; administrative tasks; conducting lab periods.

Scale: none; <10%, 10–24%, 25–49%, 50–74%, 75–100%

How often do you use the following teaching methods or media?
Lecture; use computers; use audiovisual material; have teacher-led whole group discussion; have students respond orally to questions on subject matter; have student-led whole group discussion; have students work together in cooperative groups; have students complete individual written assignments or worksheets in class; have students give oral reports.

Scale: Never, 1–2 times a month, 1–2 times a week, almost every day, every day

SOURCE: NELS Second Follow-up Teacher Survey (1992).

Although non-NCES studies have begun to collect data in a wide range of different ways (as discussed in the next section), NCES has to date relied solely on survey instruments (and largely on *teacher* surveys) as a means of collecting information on curriculum, pedagogy, and instructional resources. There are several reasons why this survey approach may be inadequate for collecting information on these areas. First, given that a respondent's time is not costless, the number of items that can be devoted to these topics is necessarily limited in multi-purpose surveys. Given the complex and multi-faceted nature of classroom activities, though, it is not clear whether sufficient useful information can be gathered in a few items, particularly on instructional practice and goals.

Second, there are serious validity issues arising in the use of general survey items. Since classroom activities vary from class to class depending on the subject, groups of students present on any one day, and student and teacher moods vary, "each section of each course results in a potentially unique content description" (Porter 1991, p. 15). Surveys are inherently static (i.e., conducted at a point in time), and continual traditional written surveying is evidently costly and impractical. But since classroom activities vary from day to day and group to group, key issues arise as to the timing of survey instruments and the reference point for teacher responses. Even if there were no variation problem, interpretation of specific activities and events is subjective. The same classroom activities from the perspectives of student, teacher, and outside observer may be very different, and NCES studies primarily rely on teacher perceptions alone.

Third, although recent broad-based national surveys such as NELS have greatly improved classroom-level information, the items tend to be descriptions of what teachers do, rather than what students do. While we may know from the handful of questions teachers are asked about curriculum content and instructional strategies that, for example, a particular classroom in 8th-grade math is lecture-based, has a certain number of units of algebra, and has two computers, these three elements could be combined in numerous ways to produce differing classroom environments and learning opportunities. Similarly, existing data tend to be based on a particular conception of the teaching process, one based on a direct-teaching, whole-classroom model. This may be inappropriate in a dynamic educational world in which many reform efforts are seeking to change important aspects of curriculum and classroom organization.

Finally, NCES surveys have collected only limited information on instructional resources, emphasizing class size, student ability, and teacher credentials. As Oakes and others have shown, school context factors can significantly affect classroom instruction and students' opportunity to learn (Guitton and Oakes 1995; Stasz et al. 1990, 1993). An important factor for interpreting survey data on the implemented curriculum and teacher practices is professional teaching conditions. If teachers receive limited staff development to enhance their repertoire of teaching techniques or learn about new approaches for teaching mathematics or other subjects, then it may not be surprising to learn that they rely on "traditional" methods in their teaching. Put another way, lack of instructional resources, including materials (e.g., computers, textbooks) and professional development, may inhibit "opportunity to teach."

IMPROVING DATA ON OPPORTUNITY TO LEARN

Up to this point, we have argued that the concept of opportunity to learn is a useful way to encompass curriculum content, pedagogy, and instructional resources in schooling. We also argue that measures of OTL can be enhanced by placing a greater emphasis on student learning, not only through tests that assess the "attained" curriculum but also by gathering information on the learning process. We further argue that while teacher survey items may provide a reasonable description of curriculum content, it is doubtful that they can ever satisfactorily assess detailed elements of classrooms without undue burden on respondents; hence other forms of data collection that are available might be considered by NCES. The goal is to be able to provide researchers and policymakers with a richer picture of what takes place within classrooms both in terms of how students learn and what students and teachers do.

In this section, we discuss various ways in which NCES data collection could be improved that draw on several recent efforts to assess OTL and that acknowledge the shift from a process-product orientation toward one that emphasizes teaching and learning for understanding. Our discussion is grouped into three sections: 1) enhancing measures of curriculum content and instructional practice items on national teacher surveys; 2) enhancing measurement of student learning processes through student surveys and other methods, such as observation, artifact collection, interviews, video data, teacher logs, and so on; and 3) enhancing instructional resource measures on surveys.

In discussing possible improvements, we draw on the contributions of several non-NCES efforts that in recent years have attempted to collect a broader array of data on opportunity to learn using a wide variety of alternate tools. These include the Reform Up Close project, conducted by the Consortium for Policy Research in Education; the Validating National Curriculum Indicators project, conducted by a team of RAND/University of California-Los Angeles researchers; and the Third International Mathematics and Science Study (TIMSS).[14] The components and data collection methods used in these efforts are summarized in Table 3. The purposes of each study differ somewhat and the methodologies used reflect this. It should be stressed that these efforts are ongoing at least in the sense that analyses of the findings are preliminary; hence any conclusions drawn from this work must be regarded as tentative. It will be several years before the data collected from these projects are fully evaluated. It seems likely, however, that they will provide valuable insights into possible enhancements to NCES data used to assess opportunity to learn.

As the name suggests, TIMSS is a cross-country study (about 50 countries in total) designed to evaluate teaching and learning in mathematics and science for 9- to 13-year-olds. It builds upon earlier similar studies (the First and Second International Mathematics Study, FIMS, and SIMS). Pilot testing for TIMSS was conducted in 1993 and 1994, and collection and analysis of data are ongoing. TIMSS includes a student achievement component, an assessment of student attitudes and background, and class-level data on opportunity to learn from teachers and school-level officials. In addition to survey data, textbooks/materials are being collected.

The RAND/UCLA study sought to assess the validity of data collected in national studies such as NELS. Its purpose was to "design and pilot a model for collecting benchmark data on school coursework" (Burstein et al. 1995, p. 1). Conducted using a small sample (just 70) of the mathematics teachers surveyed in the 1992 NELS sample of 12th-grade math teachers, the projects' preliminary findings have been published (Burstein et al. 1995). Enhanced versions of various NELS teacher survey items were administered to these teachers (see next section). The project relied heavily on the collection and coding of artifacts such as classroom assignments, quizzes and exams, and textbooks, as well as daily teacher logs and some teacher interviews (originally unplanned).

Finally, the Reform Up Close project, funded by the National Science Foundation, compiled data on secondary mathematics and science in high schools in 12 school districts, using an array of different collection strategies. Since the focus of the study was an investigation of state and local reforms in math and science, interviews were conducted with state- and-district level administrators, as well as school principals and teachers. A teacher survey was administered

Table 3—Selected recent non-NCES studies assessing opportunity to learn

	Curriculum content	Instructional practice	Instructional resources
Third International Math and Science Study (TIMSS)	Topic coverage (text, teacher survey); time on topics (teacher survey); emphasis on topics (teacher survey); topic test items (teacher survey)	Teaching practices (teacher survey); student activities (teacher survey); classroom management (teacher survey); grading and homework (teacher and student survey); planning time (teacher survey)	Teachers' knowledge of topics (teacher and student survey); texts, equipment, facilities (teacher survey)
Validating National Curriculum Indicators (RAND/UCLA) (Burstein et al. 1995)	Topic coverage (teacher logs, artifacts, text, teacher survey); time on topics (logs, artifacts, text, teacher survey); emphasis on topics (logs, artifacts, text, teacher survey); topic test items (artifacts, teacher survey)	Teaching practices (logs, artifacts, teacher survey); student activities (logs, artifacts); classroom management (logs); grading and homework (artifacts, teacher survey)	Teachers' knowledge of topics (teacher survey); texts, equipment (teacher survey)
Reform Up Close (CPRE/RUC)	Topic coverage (logs, texts, observation, interview, teacher survey); time on topics (logs, observation, teacher survey); emphasis on topics (logs, teacher survey); topic test items (logs)	Teaching practices (logs, observations, interviews, teacher survey); student activities (logs, observations, teacher survey); classroom management (observations); grading and homework (logs, teacher survey)	Teachers' knowledge of topics (interviews, teacher survey); texts, equipment (logs, observations, interviews, teacher survey)

SOURCE: Adapted by the authors from NCES (1995b), Table 1-Summary, pp. 35–38. See NCES (1995b) for full references used to compile the information in this table.

to around 400 teachers; a subset of these (between 62 and 82) were observed by the researchers; and daily activity logs were completed. "The effort to represent 'opportunity to learn' in the classroom was an important part of the larger study but not the primary focus" (NCES 1995b).

Enhancing Measures of Curriculum Content and Instructional Practice Items on National Teacher Surveys

One relatively straightforward extension of existing NCES data efforts would be a revision and extension of existing NELS-type items on curriculum content on teacher surveys. Some guide as to how this might be done is provided by the three studies cited above, in particular the RAND/UCLA study that built directly on the NELS items. Table 4 shows some typical extensions of the NELS items used by this study, based on the original teacher survey items reported in Table 2.

The curriculum content items probe for topic coverage at a greater level of specificity than previous surveys and also ask about the number of class periods teachers spend on each topic (although as few as 10–15 minutes counts as a "period" on their scale). The enhanced instructional practice items include strategies advocated in mathematics reform efforts. Burstein et al. (1995) also scaled and factor analyzed these items to see if they could meaningfully define instructional "repertoires"—instructional strategies that occur together. Such repertoires might provide a more coherent picture of instruction than simply reporting frequencies of teaching behaviors on an item-by-item basis. Although their analysis was hampered by a lack of variation in classroom practices across teachers in their sample, their approach looks promising for assessing instructional repertoires and might be used in future studies, particularly those that try to link repertoires to student outcome data.

The Burstein et al. (1995) validity study concludes that it is possible to add further, more refined topic areas to curriculum content items, and also additional questions on instructional strategies albeit with close attention to the response scale provided. The research team further recommends dropping items relating to instructional goals. Given the "paucity of empirical work regarding the definition and validation of curriculum-specific instructional constructs" (Stecher 1992, p. 76), the recommendation of Burstein et al. (1995) for future validation studies makes sense. They suggest that at the outset of large-scale national surveys, in-depth studies of small samples of teachers be conducted, using techniques that measure instructional processes with greater subtlety than is possible through survey data (Burstein et al., p. 56). These recommendations recognize that the language of instruction is in a state of flux, which may partly account for findings of lower validity for content and practices associated with the mathematics reform movement.

Earlier surveys relied on findings from process-product studies to identify "effective" practices to develop items (NCES 1995b), while more recent surveys have used state curriculum frameworks and reports from various professional groups, such as NCTM. Future studies to assess teaching practices should also look to current research on "teaching for understanding" (Good 1995); Blumenfeld's research on science teaching (1992) or research on teaching from the

Table 4—Examples of enhanced NELS teacher survey items on curriculum and pedagogy from RAND/UCLA Study

Curriculum Content

Have you taught or reviewed the following topics during this year in class?
Estimation; proportional reasoning; tables and charts; graphing; math modeling; ratios; proportions and percents; conversion among fractions decimals and percents; laws of exponents; inequalities.

Scale: No, but it was taught previously; Yes, but I reviewed it only; Yes, I taught it as new content; No, but I will teach or review it late this school year; No, topic is beyond the scope of this course.

Indicate the appropriate number of periods devoted to each topic. If you focus on topic for 10 or 15 minutes on a given day, count that as a period.
Topic list above.

Scale: None; 1 or 2 periods; 3–5 periods; 6–10 periods; more than 2 weeks, but less than 1 month (11–20 class periods); 1 month or more.

Instructional Practice

How often do you use the following instructional strategies with this class?
Demonstrate working an exercise on the board; have student work on exercises on the board; use manipulatives to demonstrate a concept; have smaller groups work on problems to find a joint solution; have students work on problems for which there is no obvious solution; have students keep a mathematics journal; have students represent and analyze relationships using tables and graphs.

Scale: Almost every day; once or twice a week; once or twice a month; once or twice a semester; never.

SOURCE: Burstein et al. (1995)

constructivist perspective (e.g., Collins et al. 1989; Stasz et al. 1993) to identify effective teaching practice. Selecting items is not straightforward, however, as it is important to be able to capture the range and variety of traditional and reform teaching practices, rather than focusing on a narrowly defined vision of practice (Smithson et al. 1995).

Enhancing Measurement of Student Learning Processes Through Student Surveys and Other Methods

An obvious way to extend assessment of curriculum and pedagogy, and to include information about student learning, is to administer student surveys. One recent attempt to administer a student survey may be found in the Smithson et al. (1995) study of middle-level science. The purpose of the survey is to assess the enacted curriculum in science to help states participating in the project to assess OTL, interpret the results, and improve classroom practice. Students are asked about the frequency with which they experience 27 separate activities (see Table 5) and about their previous exposure to science. Teachers of these students are asked about their educational background, influences on the curriculum of the science class, how computers and calculators are used, and homework and grading policies.

Most interestingly, teachers are also asked about the instructional activities of the student that are directly aligned to the 27 student items: How often does an average student do these things in science class? Thus, the responses of teachers and students can be compared. Items for teachers and students emphasize three kinds of activities: acquiring information, using information, and extending information. Since questionnaires were field-tested in the 1994–95 school year, results from this study will be forthcoming. Their approach, however, looks promising for assessing student learning and its relationship to teaching practice and student outcomes.

Survey data have been the only type of data NCES has sought to collect on a national basis on opportunity to learn. Other researchers have utilized a wide variety of methods, and in recent years these have been tied to survey data on curriculum and pedagogy either as a means of determining the validity and reliability of survey items or as a research tool in their own right. These alternative forms include the following: collection of teacher lesson plans; written assignments, exams, and textbooks; teacher logs detailing classroom time allocation and tasks; observation and video of classrooms in action; and teacher interviews.

Burstein and his colleagues (Burstein et al. 1995), for example, gathered teacher assignments (homework, quizzes, classroom exercises, projects, examinations) as a way to validate teacher reports of their practices because they represent much of the curriculum presented to students. These items probe the types of performance teachers expect from students; for example, what percent of test items "require a critique or analysis of a suggested solution to a problem" or "require the application of concepts and principles to different or unfamiliar situations." Similarly they ask how frequently teachers assign various types of homework, such as "gathering data, conducting experiments, working on projects" or "explaining newspaper/magazine articles."

In keeping with the primary purpose of this study, the authors compare test or homework survey items to artifacts (actual tests and assignments) to assess the validity of survey responses. In both cases, agreement was low, suggesting that surveys are not very reliable for assessing what teachers expect of students, particularly for "more innovative" items that reflect a reform-oriented perspective (i.e., encourage student-centered activity and construction of knowledge). They also recognized, however, that the curriculum presented through these artifacts does not provide information about how students receive and respond to the curriculum. To get a sense of student

Table 5—Examples of student survey items on opportunity to learn

How often do you do the following activities in your science class?

Listen to your teacher or someone else explain things about science
Read about science in books, magazines, or articles in class
Collect data from sources in books, magazines, or articles in class
Read tables, graphs, or charts
Use measuring tool such as rulers, thermometers, balances, computers, and so on
Do a laboratory activity, investigation, or experiment
Observe experiments or investigations that others do, including teacher demonstrations
Watch films or videos
Use laboratory equipment
Work in small groups
Participate in school planned and supervised activities outside the classroom
Work on assigned science projects or activities on your own away from school
Use the computer in science
Answer questions from your science book
Take a quiz or test
Write about science (e.g., lab reports, science papers)
Make your own tables, graphs, and charts
Change something in an experiment to see its effects
Design experiments
Ask questions to improve your understanding
Make predictions, guesses, or hypotheses
Make maps/drawings or models to show scientific ideas
Reach conclusions about scientific data
Choose a method for expressing an idea to your class
Revise and improve your work
Apply scientific concepts to your everyday life
Explain what you learn in science relates to real-world issues (such as the environment)

Scale: Nearly every period; about once a week; once or twice a month; once or twice a year; never.

Teachers are asked about same list: ***How often does an average student do these things in science class?***

SOURCE: Smithson et al. (1995). Items taken from middle-level science student survey.

learning, they asked teachers to provide examples of student work associated with each major assignment. This request appeared overly burdensome for teachers, however, and this data collection was subsequently abandoned (Burstein et al. 1995).

Teacher logs were collected by both RUC and RAND/UCLA studies as a way of mapping out the daily activities of teachers. The RAND/UCLA study was designed so that researchers were able to directly compare instructional practice survey items (such as those in Table 4) with teacher logs and artifacts such as homework assignments, quizzes and exams, and textbooks. The teacher logs were completed at the end of each day, collecting information on topic coverage, student activities, and modes of instruction. In keeping with the small-scale "benchmarking" purposes of the study, logs and artifacts were collected only during a 5-week period. The rate of agreement between surveys and logs was "quite low," although part of the explanation for this finding "may lie in how the survey response categories were constructed" (Burstein et al. 1995, p. 39).

RUC used one-page daily logs in which teachers recorded lesson topic, subtopic, presentation mode, and student performance expectations using a 4-digit code; this permitted the coding of, for example, almost 6,000 different 4-digit content characterizations of math lessons (NCES 1995b). These data were collected from teachers over an entire year, and compared for some lessons with structured observer reports focusing on instructional activities, student engagement, and classroom management. While comparison of survey data with observations and logs showed only a moderate degree of agreement, observers' reports and teacher daily logs for the same lessons "indicated that some dimensions of instruction could be described with a high degree of inter-rater reliability in an activity that takes only a few minutes a day" (NCES 1995b).

While surveys, artifact collection, teacher logs, and teacher interviews can provide a great deal of useful information about curriculum content and teacher activities that occur frequently and are well established, "some aspects of curriculum practice simply cannot be measured without actually going into the classroom and observing the interactions between a teacher and students" (McDonnell 1995, p. 310). This truth raises fundamental issues for national data collection efforts if richer data on classrooms are to be collected. Clearly there are limits to the extent to which outside observers can enter classrooms and assess lesson content, instructional strategies, and student activities, both in terms of cost and in terms of generating useful information. Observing lessons is labor intensive and hence very costly if done on a large scale. Using observations in validity studies, i.e., to check the response to survey data, may be confined to small samples, but even here there is a concern about the representative nature of the subsample of teachers observed. There are difficulties in deciding which classrooms to pick and when to observe them. This problem was encountered to a degree in the Burstein et al. (1995) study in which daily logs and artifacts were collected from just 70 teachers. It turned out that the background qualifications in mathematics of these teachers was considerably different from the wider NELS sample, raising doubts about the study's overall findings.

Further, if data gathered through these non-survey methods are to be used for purposes beyond simply assessing the validity of survey items, they would ideally be generalizable to some extent. This implies the need for structured forms to record classroom observations and careful observer training so that similar behaviors are recorded as similar by different observers. RUC utilized a structured observers' form as a means of assessing log and survey items, but also offered them an opportunity to provide a narrative report dealing with more subtle aspects of what was taking place in the classroom. It is not clear that methodological techniques exist to the point where such data could be coded or analyzed in ways that produce generalizable findings at reasonable cost.

One possible way in which the costs of observation and difficulties of interpretation may be reduced is through the collection of video data. The TIMSS Videotape Classroom Study currently under way is collecting information about classroom mathematics instruction to supplement data from assessments and questionnaires collected in the main TIMSS study (Stigler and Fernandez 1995; see also the paper by James Stigler for this NCES Futures Project). The project collected a random sample of approximately 100 TIMSS 8th-grade classrooms in the United States and Germany, and 50 in Japan. The tapes have been transcribed onto CD-ROM and linked in a multimedia database to translated transcripts of classroom speech in order to enable computer access to video data. Tapes will be coded to describe classroom instruction in the three countries and can be linked to survey data on classroom instructional methods. Stigler and Fernandez (1995) describe field-test study procedures and lessons learned thus far in collecting video data (including hiring and training videographers) and designing the multimedia database. The outcomes of this project will have important implications for judging the costs and feasibility of using video data to track and assess teaching practice, student process, and so on.

Enhancing Instructional Resource Measures Using Surveys

The non-NCES studies discussed here as well as others point to an inconsistency between the rhetoric of reform movements and the reality of teaching practice. For example, Burstein et al. (1995) found internal consistencies in teacher surveys on reporting instructional practices and goals. However, follow-up interviews with teachers revealed that teachers did not know what "math modeling" meant, even though it appears in the state math curriculum frameworks. The authors conclude that survey data may not be validly interpretable at a time when practice is in flux. Their findings also suggest that additional information about instructional resources—particularly teacher professional development—might improve our ability to interpret survey data. As discussed earlier, if teachers lack staff development opportunities or work in a school where the community of practice does not support *teacher* learning, then how are teachers to come to understand "transitional" curriculum content and instructional practices? At the very least, additional survey items might assess "opportunity to teach."

Current surveys rely on teacher background characteristics as indicative of teacher quality and ask teachers to report degrees and credentials, undergraduate major, the subjects they teach, and the like. These indicators tie "quality" to knowledge of subject matter and credentials. In addition to this information, surveys might assess teachers' knowledge of and opportunities to learn about innovations, with questions like the following: How many workshops or other professional development activities have you attended this year? How many focused on new curriculum and instructional practices in mathematics? How would you rate the usefulness of these activities . . . for improving understanding of mathematics teaching reforms? . . . for changing the way you teach mathematics? . . . for changing the kinds of assignments you give to students? How often are you able to discuss new ideas about mathematics teaching with other teachers at your school? Does your school and district support teachers who want to adopt innovative curriculum and teaching practices? (Also, see Smithson et al. 1995 for a teacher survey that includes questions about professional development opportunities.) Broadening teacher background measures in this way would help strengthen existing data on instructional resources.

CONCLUSIONS AND RECOMMENDATIONS

In this paper, we have used the concept of opportunity to learn to frame our discussion of national data gathering efforts on curriculum and pedagogy. We have discussed the purposes of data collection in this area, reviewed NCES data collection based on national surveys, and drawn on some very recent studies that have refined survey items and widely utilized non-survey modes of data collection. Given the preceding discussion, we conclude with a set of specific recommendations that NCES might adopt with regard to improving data on opportunity to learn.

First, as demonstrated by recent non-NCES studies, teacher surveys seem an effective and efficient way to gather information about course taking if the standard is knowing whether or not a topic has been taught, and if it has been taught over several periods or weeks. More reliable data can be obtained by asking teachers more specific questions about particular curricular topics. Adding more finely grained items to teacher surveys would thus appear to be sufficient for gathering national data on curriculum content. More instructional strategy questions could be added to teacher surveys; in particular, to expand the types of teaching practices to assess any shifts from traditional to "reform-oriented" pedagogy. (Research from the "teaching for understanding" paradigm provides a source for identifying new items.) However, while practice is in transition, it appears to be worth continuing validity studies. This way, future efforts can provide reliable estimates of changing teaching practice. Collection of non-survey data (such as teacher daily logs, teacher interviews, and observations), and use of exams, quizzes, assignments, and textbooks for small subsamples of survey respondents apparently provide useful additional information that help supply a richer picture of classroom practice and a means of assessing the validity of survey items. Methods of coding such data have been developed in recent years (for example, by the Burstein et al. team). The feasibility of using videotape data such as that collected by the TIMSS project should be closely monitored by NCES, and a similar effort should be considered in conjunction with a future national data collection such as the Early Childhood Longitudinal Study.

Second, enhanced efforts to assess student learning should be undertaken in future NCES work. Although newer studies have attempted to include items about teaching practice that reflect the constructivist view, they remain teacher-oriented. Adding a stronger student component might be accomplished by adding class-specific items to student surveys that mirror those on teacher surveys, perhaps with classroom observations to assess item validity. Further, the enriched data collected by non-survey methods, such as collection of artifacts and video data, may be utilized in aiding understanding of student learning. We recommend that NCES begin to explore the usefulness of these methods for collecting information of student learning processes as a supplement to survey data.

Third, further items on teachers' background and preparation for teaching—opportunity to teach—could also reinforce existing teacher surveys. Since many curricula and pedagogical reforms place new demands on teachers, it is important to determine whether or not teachers are equipped to adopt and effectively utilize these new methods. Currently, measures on instructional resources at the classroom level are limited in their number and scope.

Despite its flaws, NCES has been a valuable asset to policymakers and researchers in helping to understand opportunity to learn and outcomes of schooling. Recent efforts in Congress propose to do away with the Department of Education and give more control of federal dollars to the states. The shift to block grants to states can only enhance the diversity we already see in our nation's schools, per-pupil expenditures, teacher student ratios, and the student learning and teaching processes that take place within classrooms. Similarly, as more states embrace "charter schools," schemes that privatize schooling-related services (e.g., Edison, Educational Alternatives Inc.), or adopt school "choice," the educational landscape promises to become even more diverse. If public education remains a federal policy issue—and we think it must—then federal efforts that gather systematic, representative data on a myriad of schooling interventions seem more vital than ever. NCES has a key role to play in this effort.

NOTES

1. The concept of OTL was first introduced in the 1960s. For a full discussion of the origins, evolving definition, and policy applications of OTL, see McDonnell (1995) and other contributions to the special issue of *Educational Evaluation and Policy Analysis* dedicated to the late Leigh Burstein 17(3), Fall 1995.

2. The distinction is often drawn between the *intended* curriculum, the *implemented* curriculum, and the *attained* curriculum (Stecher 1992; McDonnell 1995). Here we are concerned with the first two of these.

3. The three reasons for collecting data on opportunity to learn distinguished below are similar to those outlined by Porter (1991).

4. For a review of process-product research, see Brophy and Good (1986) and Good (1995). Some recent work on curriculum and student achievement is also noted below.

5. For a review of "effective schools" literature, see Purkey and Smith (1983) and Rosenholtz (1985). For a useful critique see Rowan et al. (1983).

6. For a comprehensive overview of this work, see Hanushek (1986) and the critique by Hedges, Laine, and Greenwald (1994). Hanushek (1979) contains a useful discussion of methodological problems associated with this type of research. More recent examples using NCES data include Ehrenberg and Brewer (1994) (HS&B) and Ehrenberg, Goldhaber, and Brewer (1995) (NELS).

7. A recent example is Kuppermintz et al. (1995). While the primary purpose of this paper is to assess the validity of NELS mathematics test items, they do find strong effects of course program indicators on achievement. For example, higher scores in mathematics knowledge are associated with teacher reports of more traditional instruction methods (p. 545).

8. One shortcoming of surveys like NAVE for tracking reform concerns timing of data collection relative to the time it may take to implement reform efforts, especially curriculum or teaching practice reforms that can take many years to put into practice (Grubb and Stasz 1992).

9. For example, the 1965 Equality of Educational Opportunity Survey (Coleman Report) teacher survey contained several questions related to teachers' qualifications and academic backgrounds, but very few directly to their classrooms (basically limited to number of students, hours spent preparing for class, number of classes taught, and attitudes toward racial issues and ability grouping). A recent re-examination of these data may be found in Ehrenberg and Brewer (1995).

10. For an overview of NCES surveys and future plans, see NCES (1995a).

11. While HS&B contains longitudinal student information, including standardized test scores at two points in time for the sophomore cohort of 1980, there is no direct link between the teachers surveyed and the students, or between the tests and particular classroom subject content. SASS includes no student component. NAEP, while containing directly linked student and teacher information on curriculum and instruction, and student standardized achievement measures, consists of representative samples of several grade/age levels; it is not longitudinal in the sense that the same students are surveyed from year to year.

12. Various "transcript studies" have been conducted in conjunction with HS&B (1982), NAEP (1987, 1990), and NELS (1992). These involve the coding and descriptive analysis of a large number of school transcripts obtained for a subsample of students in the relevant national survey (NCES 1995). They provide additional information about student course-taking patterns, though they do not indicate detailed curriculum content.

13. For details on future surveys, see NCES (1995a). The major scheduled vehicles for future data collection on curriculum and pedagogy at the K–12 will be NAEP (continuing every 2 years), and the Early Childhood Longitudinal Study (ECLS), planned to begin pilot testing in 1996–97. SASS (continuing every 5 years) will collect teacher data.

14. More details on these studies as they relate to "opportunity to learn" may be found in NCES (1995b).

REFERENCES

Blank, R. K. 1993. "Developing a System of Education Indicators: Selecting, Implementing, and Reporting Indicators." *Educational Evaluation and Policy Analysis* 15 (1): 65–80.

Blumenfeld, P. 1992. "The Task and the Teacher: Enhancing Student Thoughtfulness in Science." In J. Brophy (ed.), *Advances in Research on Teaching* 3: 81–114. Greenwich: JAI Press.

Brophy, J., and Good, T. 1986. "Teacher Behavior and Student Achievement," In M. Whittrock (ed.), *Handbook of Research on Teaching*. New York: MacMillan.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., and Guitton, G. 1995. DRU-1086-NSF. *Validating National Curriculum Indicators*. Santa Monica, CA: RAND.

Collins, A. In press. "Design Issues for Learning Environments." In S. Vosniadou, E. DeCorte, R. Glaser, and H. Mandl (eds.), *International Perspectives on the Psychological Foundations for Technology-Based Learning Environments*. Hillsdale, N.J.: Erlbaum.

Collins, A., Seely, J., and Newman, S. 1989. "Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing and Mathematics." In L. B. Resnick (ed.), *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*. Hillsdale, N.J.: Erlbaum.

Ehrenberg, R. G., and Brewer. D. J. 1994. "Do School and Teacher Characteristics Matter? Evidence from High School and Beyond." *Economics of Education Review* 13 (1): 1–17.

Ehrenberg, R. G., and Brewer, D. J. 1995. "Did Teachers' Verbal Ability and Race Matter in the 1960s? *Coleman* Revisited." *Economics of Education Review* 14 (1): 1–23.

Ehrenberg, R. G., Goldhaber, D. D., and Brewer, D. J. 1995. "Do Teachers' Race, Gender, and Ethnicity Matter? Evidence from NELS:88," *Industrial and Labor Relations Review* 48 (3): 547–561.

Erickson, F., and Shultz, J. 1992. "Students' Experience of the Curriculum." In P. W. Jackson (ed.), *Handbook of Research on Curriculum*. New York: MacMillan.

Good, T. L. 1994. "Teaching Effects and Teacher Evaluation." Draft manuscript of a chapter to appear in the second edition of the *Handbook of Research on Teacher Education*.

Grubb, W. N., and Stasz, C. 1992. *Assessing the Integration of Academic and Vocational Education: Methods and Questions* (MDS-445). Berkeley: National Center for Research in Vocational Education.

Guitton, G., and Oakes, J. 1995. "Opportunity to Learn and Conceptions of Educational Equality." *Educational Evaluation and Policy Analysis* 17 (3): 323–336.

Hanushek, E. A. 1979. "Conceptual and Empirical Issues in the Estimation of Education Production Functions." *Journal of Human Resources* 14 (3): 351–388.

Hanushek, E. A. 1986. "The Economics of Schooling: Production and Efficiency in the Public Schools." *Journal of Economic Literature* XXIV (3): 1141–78.

Hedges L., Laine, R., and Greenwald, R. 1994. "A Meta Analysis of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (3): 5–14.

Hoffer, T. B., Rasinski. K. A., and Moore, W. 1995. *Social Background Differences in High School Mathematics and Science Coursetaking and Achievement* (NCES 95-206). Washington, D.C: U.S. Department of Education.

Jones, L. V., Davenport, E. C., Bryson, A., Bekhuis, T., and Zwick, E. 1986. "Mathematics and Science Test Scores as Related to Courses Taken in High School and Other Factors." *Journal of Educational Measurement* 23 (3): 197–208.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., and Snow, R. E. 1995. "Enhancing the Validity and Usefulness of Large-Scale Educational Assessments: NELS:88 Mathematics Achievement." *American Educational Research Journal* 32 (3): 525–554.

Kozol, J. 1991. *Savage Inequalities: Children in America's Schools.* New York: Crown.

McDonnell, L. M. 1995. "Opportunity-to-Learn as a Research Concept and a Policy Instrument." *Educational Evaluation and Policy Analysis* 7 (3): 305–322.

Monk, D. H. 1992. "Educational Productivity Research: An Update and Assessment of its Role in Education Finance Reform." *Educational Evaluation and Policy Analysis* 14: 307–332.

Murnane, R. J., and Phillips, B. 1981. "What do Effective Teachers of Inner-City Children Have in Common?" *Social Science Research* 10: 83–100.

National Center for Education Statistics (NCES). 1995a. *Programs and Plans of the National Center for Education Statistics*. Washington, D.C.: U.S. Department of Education.

National Center for Education Statistics (NCES). 1995b. *Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work* (NCES Working Paper 95-11). Washington, D.C.: U.S. Department of Education.

National Council of Teachers of Mathematics (NCTM). 1991. *Professional Standards for Teaching Mathematics*. Reston, VA: NCTM.

Oakes J., Selvin, M., Karoly, L., and Guitton, G. 1992. *Educational Matchmaking: Academic and Vocational Tracking in Comprehensive High Schools* (MDS-127). Berkeley: National Center for Research in Vocational Education.

O'Day, J., and Smith, M. S. 1993. "Systemic Reform and Educational Opportunity." In Susan Fuhrman (ed.), *Designing Coherent Education Policy*. San Francisco: Jossey-Bass.

Porter, A. C. 1991. "Creating a System of School Process Indicators." *Educational Evaluation and Policy Analysis* 13 (1): 13–29.

Porter, A. C. 1995. "The Uses and Misuses of Opportunity to Learn Standards," *Educational Researcher* 24 (1): 21–27.

Purkey, S. C., and Smith, M. S. 1985. "Effective Schools: A Review." *The Elementary School Journal* 83 (4): 427–452.

Reigeluth, C. M. 1987. *Instructional Theories in Action: Lessons Illustrating Selected Theories and Models*. Hillsdale, N.J.: Erlbaum.

Rowan, B., Bossert, S., and Dwyer, D. 1983. "Research on Effective Schools: A Cautionary Note." *Educational Researcher* 4: 21–31.

Serrano v. Priest. 1971. 487 P.2d 1241. "*Serrano v. Priest*: Implications for Educational Equality." *Harvard Educational Review* 41 (4): 501–534.

Shavelson, R., McDonnell, L. M., Oakes, J., and Carey, N. 1987. *Indicator Systems for Monitoring Mathematics and Science Education* (R-3570-NSF). Santa Monica: RAND.

Smithson, J. L., Porter, A. C., and Blank, R. K. 1995. *Describing the Enacted Curriculum: Development and Dissemination of Opportunity to Learn Indicators in Science Education*. Washington, D.C.: Council of Chief State School Officers.

Stasz, C., McArthur, D., Lewis, M., and Ramsey, K. 1990. *Teaching and Learning Generic Skills for the Workplace* (R-4004-NCRVE/UCB). Santa Monica: RAND.

Stasz, C., Ramsey, K., Eden, R., DeVanzo, J., Farris, H., and Lewis, M. 1993. *Classrooms That Work: Teaching Generic Skills in Academic and Vocational Settings* (MR-169-NCRVE/UCB). Santa Monica: RAND.

Stecher, B. 1992. *Describing Secondary Curriculum in Mathematics and Science: Current Status and Future Indicators* (N-3406-NSF). Santa Monica: RAND.

Stigler, J., and Fernandez, K. 1995. "Videotape Classroom Study: Field Test Report" (unpublished technical report). Los Angeles: University of California.

Teacher Education, Training, and Staff Development: Implications for National Surveys

David R. Mandel

INTRODUCTION

The national conversation about teaching has always been compromised by a dearth of information about the quality of practice and practitioners. Without such information, policymakers and the public have been left to fend for themselves, often speculating wildly about the current state of teacher education and teaching. When dismal or promising results about student performance are reported, a new chain reaction of suppositions is often set off about the degree to which teachers are to be blamed or praised. But these suppositions are just that—hypotheses disconnected from much of a factual base that might shed some light on what is occurring, including the extent to which the observed results can be accurately attributed to teacher actions.

All of this should not be the least bit surprising, because until recently the profession had not been able to agree on just what are the essential elements of highly accomplished practice. Without such a definition, those charged with collecting data on teacher quality were left in the lurch. As a result, to the extent that any qualitative measures of teacher competence exist, they are marginal at best. This dilemma also hobbles the necessary debate that should be ongoing about teaching, and it compromises the ability to conduct useful research and analyses about teacher, school, and system performance. These circumstances have led to the growth of some familiar myths about teaching, such as all that matters is teachers' command of subject matter knowledge; teachers just need to stay one chapter ahead of students and employ the latest pedagogical trick; or anyone who is a warm, nurturing, and caring individual can be a teacher. By abdicating its responsibility to put forth a rich description of excellent practice, the profession has contributed to the promulgation of these notions that teaching is for amateurs, not professionals, and that anyone with a good heart and a modicum of intelligence stands a decent chance of being successful.

A UNIQUE OPPORTUNITY TO ADDRESS THE TEACHER QUALITY ISSUE

Fortunately, the nation is moving to a new reality as, for the first time, a systematic and prolonged effort is being undertaken to develop in each field of teaching a professional consensus about what accomplished teachers should know and be able to do. Constructed as part of a larger effort by the National Board for Professional Teaching Standards (NBPTS) to create a voluntary system of advanced professional certification for early childhood, elementary, middle, and secondary school teachers, this new reality has emerged in the form of standards for advanced practice that provide a foundation for important conversations about teacher quality that previously could not be conducted (NBPTS 1994, 1995, and 1996).

The National Board standards, focused as they are on the critical aspects of teaching that distinguish the practice of accomplished professionals, create a framework for evaluating teachers, institutions, and programs that are designed to develop exemplary teachers. Such a framework heretofore could not be provided by teacher licensing standards, for important as they continue to be to the states (as they exercise their responsibility to assure basic competence in individuals assigned responsibility to educate the young and vulnerable), they are constrained by the function they are designed to serve. Concurrently, these standards for exemplary practice allow the profession to claim the legitimacy it is due, because they define the expertise and best practices that not only distinguish highly accomplished practitioners from beginners and journeymen, but also make clear to the public as well as professionals the elegance that marks excellent practice when the art and science of teaching are well joined.

The National Board's vision of accomplished teaching recognizes that the world for which we are preparing our students is importantly different from the mission schools have focused on in the past. Like the best of the recent student standards initiatives, this view of teaching is built on a common ground of assumptions that begin with an acknowledgment that the nation can no longer afford to provide an excellent education to a small elite and a pedestrian education to the masses. The reasons are not only our changing economic circumstances (which are quite significant in and of themselves), but also the need to have a much better educated citizenry to sustain the nation's democratic values and institutions. These standards efforts also proceed from the shared view that neither civic competence nor labor market success can be achieved by simply stuffing students full of facts, rules, and theorems. While such a knowledge base is important, it must be augmented with the ideas, concepts, theories, and knowledge of the core disciplines that allow the well-educated to constructively address the challenges that face us daily at home, in our communities, and in the workplace. The ideas that teaching should be more than telling, that students should be actively engaged in learning and applying knowledge, that in-depth understanding is to be valued over coverage, and that all children can learn are at the heart of this perspective.

This new and emerging agreement on first principles captured in the policies and standards of the National Board for Professional Teaching Standards (NBPTS 1991) provides an opening for the National Center for Education Statistics (NCES) to aggressively pursue a set of issues that have interested it for some time, but up to now have proven elusive. As one reviews report after report about teaching that NCES has commissioned in the aftermath of *A Nation At Risk*, one is struck with the regularity with which the authors stub their collective toes on the teacher quality issue—not for lack of trying, but because there has been little data with which to work.

In the past several years, the Center has made important strides in uncovering the complicated world of teacher migration, helping the policy community better appreciate the intricacies of teacher movements from position to position within the system and of entry into and exit from the system. This has illuminated the discussion of teacher supply and demand, probably shattered some old myths that needed to be set aside, and made the discussion of this subject much more sophisticated and valuable. Still, having a better grasp of the raw flows of teachers (potential, current, and former) in and around schools provides modest solace when there remains little good understanding of the qualitative dimensions of such flows (e.g., are the teachers leaving the work force stronger than those staying, or vice versa). If there is an increase or decrease in new teachers entering or mature teachers leaving the profession should the nation be concerned? The answer depends on the mix of teachers involved and the quality of the stream of those entering, migrating, and leaving—on this score parents, policymakers, and administrators are largely in the dark.

What is known is the type of education credentials teachers have accumulated and the type of state licenses they have been granted. This information has proven useful in gaining a rough sense of how well-prepared teachers are to take on the assignments they are handed; thus, the public has been alerted to the high percentage of students whose education has been entrusted to individuals who do not have a basic grounding in the subject(s) they teach. But such data, even when positive, provide only the most modest threshold of confidence regarding the quality of practice in the nation's schools. This is so for several reasons, not the least of which is that licensure requirements differ markedly from state to state,[1] and the fact of completing a major or minor in a subject holds substantively different degrees of meaning from campus to campus. However, even if these issues of non-uniformity could be swept off the table in a stroke, yet another fundamental problem remains: assurances of minimal competence tell us hardly anything about the quality of practice once teachers have moved beyond their first few years of practice. This is a field where expertise takes time to develop and where exemplary teachers build a repertoire of content-specific pedagogy over time on which they can draw, refine their ability to understand what their students grasp about important matters, and become more astute in making the sound and principled professional judgments about how best to proceed on a daily basis—decisions that are crucial if all students are to be presented with the requisite opportunities to learn to their fullest. In short, only the first phase in educating and developing accomplished practitioners has been completed when a first degree is awarded and a first license granted. Thus, if the education community and the public at large are to have a genuine sense of the quality of the teacher work force, data must be collected on a regular basis from across the profession that are representative of the various career paths and education and work regimens that teachers have experienced and that have a profound influence on the quality of teaching in America's schools. Simply put, we need to move toward gathering information on what teachers know and can do today, and to adopt an analytical stance that is less fixated on the knowledge, skills, and dispositions they brought with them on their first day of teaching.

THE CENTRALITY OF INDICATORS OF TEACHER QUALITY TO EDUCATIONAL POLICY

This paper focuses on teacher quality because it is the quicksand that mires almost all large-scale data analysis of the state of teaching, teacher education, training, and staff development. Although NCES collects useful information on many characteristics of the nation's teachers and the type of educational experiences to which they have been exposed, the value of such data is diluted and the investment in its collection marginalized when the necessary qualitative information that ought to be linked to these other characteristics is absent. The result is that with only the most perfunctory descriptive information collected and reported, policymakers can and do fail to take critical actions that they might otherwise take to bolster the quality of education our children receive, or act precipitously when such actions are not warranted, or both.

The large policy issues that should drive data collection and analyses in this arena can be captured in the following quality assurance themes:

- The overall quality of the teaching force;
- The quality of the teacher education system; and
- Trends in teacher quality, institutional supports, and incentives for improvement.

The Overall Quality of the Teaching Force

The bedrock of the nation's schools are the teachers who are its front-line workers. The schools cannot be any better than their teachers, who, if they are to move students to high levels of performance, must make careful professional judgments on a daily basis about how students spend their time and on what subjects, ideas and concepts they focus their attention. Teachers create or fail to create learning environments that motivate student effort and stress democratic values. They teach or fail to teach the perspectives, dispositions, forms of inquiry, and ways of knowing that mark the core disciplines and that provide the pathways to the important ideas that students need to grasp. They may increase to some greater or lesser degree students' ability to learn on their own, to work collaboratively with others, and to develop the values, character, and knowledge that will allow them to function well as adults in the marketplace and in their communities and its democratic institutions. Just what the capacity of the teaching force is to perform these functions and the extent to which they are performed well is crucial information that is desperately needed from both a public policy and public administration perspective. Data are needed not only on an aggregate basis but also with respect to particular teaching fields and locales. Without such knowledge of the teaching force, those seeking to improve the schools to meet the very real challenges this society faces both externally and internally are operating with one hand tied behind their back, ignorant of the depth and breadth of the problems they confront or the assets they possess at a key point of leverage in the system.

The Quality of the Teacher Education System

A key determinant of the quality of teaching is the kind and quality of investments that are made in the initial and continuing education of teachers throughout their careers. Here, too, there is the thinnest veneer of hard evidence on which to draw conclusions, yet one can find both an extraordinary degree of complacency in some quarters about the state of this substantial enterprise and calls for the complete dismantling of the system in others. That both views have legitimacy concurrently is quite stunning. It is understandable, however, once one recognizes that in the absence of reliable and trustworthy measures (e.g., a national accreditation system with teeth to which all states subscribed, or a teacher education "NAEP") to provide a bulwark against unsubstantiated claims, such claims, no matter what distance from reality, can gain currency.

This circumstance is highly problematic in the current policy environment where for the past decade teacher education, be it undergraduate, graduate, or continuing professional education, has been under attack by its clients, the policy community, and many of its most distinguished members. Some of the concerns appear ill-founded, whereas others appear well-grounded. While it is true that accreditation by the National Council for the Accreditation of Teacher Education (NCATE) provides an increasing measure of quality assurance for those schools of education that do open themselves to critical examination by their peers, most do not. Furthermore, no parallel check on quality exists for the millions of dollars sunk into the ongoing professional development of teachers that occurs outside NCATE institutions. Even within NCATE schools, all that is known is that a basic threshold of quality has been met. The degree to which these institutions in general or the particular programs that they operate exceed the threshold remains a mystery.

Trends in Teacher Quality, Institutional Supports, and Incentives for Improvement

Although educational institutions are often characterized by their insularity to externally imposed change and by a high level of inertia even in response to internally provoked change, they have their dynamic aspects. Similarly, the teaching force itself rarely remains static for long, as shifts in the demography of the country and in the particulars of the labor market inexorably make their presence felt over time. Even without such forces at play, changes in the number of students, pupil-teacher ratios, compensation, and other conditions of work often yield effects of similar character. These shifts may be unevenly felt in different regions of the country and in different teaching fields, but they are unavoidable. NCES has played a valuable role in helping the education system gauge well the nature of such trends in the past and should continue to do so in the future.

To date, however, the focus has been on large flows through the system without much attention to quality for all of the reasons cited above. With the addition of some qualitative dimension to such analyses much more informative work could be conducted. More robust data on teacher quality would also allow analysts to look at the effects of various incentives and programmatic initiatives to improve teaching that are now, for the most part, elusive.

GOVERNING PRINCIPLES

In thinking through a new strategy for NCES to consider in this arena, a few key maxims might serve to guide the conversation. For example:

- "You can't boil the ocean." This rule, borrowed from Lewis Branscomb when he was Chief Scientist at IBM in the 1980s, is the obverse of the "there is no silver bullet" theory. It warns that agencies need to think hard about what investments are likely to have the greatest payoffs, as it is too easy to fall into the trap of trying to address every known issue to the detriment of addressing a few critical ones well.
- The easiest data to collect may not be especially valuable. Knowing how many teachers will retire this year has modest value until that information can be placed in some larger context and some further information gathered about just whom is

retiring (e.g., what they teach and whether or not they are among the strongest teachers in the system).

- Some cheap proxies may be worth looking at, either because they are cheap or because NCES has a long history of collecting them and their behavior is telling in one way or another.
- Find out what the customers value (in the current plan and prospectively) by talking to federal, state, and local policymakers; teacher educators; the unions; disciplinary and specialty associations; and the research community.
- Qualitative indicators in this field that are likely to be trustworthy are not going to emerge from counting anything (at least for the next several years, at which time the percentage of National Board Certified Teachers in a jurisdiction will mean something). Such indicators depend on the professional judgments of well-educated and well-trained examiners/observers.

CRITICAL PERSPECTIVES

In thinking through an agenda for future NCES initiatives that are focused on teachers and teaching, the history and policy questions sketched out above provide the backdrop for decision making. In considering the range of options that seem most promising to explore in light of these factors, four core frames of reference are suggested:

- *Qualitative measures of individuals*—which are without question the most important measures to capture;
- *Qualitative measures of institutions*—which pose substantial conceptual, economic, and political challenges, but would represent a breakthrough of significant proportions in understanding teaching at the postsecondary level;
- *Flows through the system*—with special attention to activities that hold the potential to promote the upgrading of the teaching force and minority participation; and
- *Kind and level of investment in teacher quality*—in time and fiscal resources by all the various players, including the teachers themselves.

A discussion of each core frame of reference follows.

Qualitative Measures of Individuals

Coming to grips with measuring the quality of teachers is not going to be addressed with a better survey instrument, even if it had an examination imbedded within it. This is because what teachers seem to know and can write about is not always an accurate predictor of what actually transpires in their classrooms. This truism has been rediscovered anew by researchers visiting the classrooms of mathematics teachers who claimed they were implementing the new National Council of Mathematics Teachers' standards. While many of these teachers spoke convincingly about how they had adapted their practice to the new standards, the actuality was often significantly different.

The hard reality is that there is no substitute for actually observing teachers at work with students. This observation could take a variety of forms, including the use of video, but it cannot be approximated by absenting the observer from the hurly-burly of the classroom. This fixation with observing the act of teaching as a necessary prerequisite to judge teacher quality is not to discount the value of NCES also providing accurate data on teachers' knowledge of the subjects they teach, of the content-specific pedagogy associated with these subjects, and of the misconceptions and difficulties about important ideas that are common among students in the age range they are teaching. As NCES proceeds down this path, it should strive to capture both what teachers know and what they can do in terms of the profession's highest standards, those established for National Board Certification in each field of teaching.

This is a tall order and presents many logistical, administrative, and economic hurdles, but with some imagination they ought to be surmountable. For starters, NCES should consider sampling teaching in a few core fields (e.g., primary grades instruction, middle grades English, high school mathematics) once every 5 years, thus providing the nation with the equivalent of a NAEP for teachers. Gaining a sense of teachers' command of the knowledge base of their field would require NCES administering some form of written examination to a sample of the teaching force, a nontrivial exercise in political and economic terms until such time as the states begin to require concrete evidence that teachers are keeping abreast of new developments in their field as part of state relicensing requirements.

More plausible in the near term would be an effort to ascertain the nature and quality of actual teaching practice across the land. To do so, the Center could take advantage of the assessment technology now being developed by NBPTS, which joins videos of teachers' practice with student work samples and teacher commentaries on the goals of instruction, the context in which the instruction is being conducted, the quality of work and understanding of the students, and teachers' rationale for proceeding as they do. These videos and their surrounding artifacts and explanations are each focused on a central responsibility that teachers in each field must discharge to advance student learning (e.g., in English language arts there is a video exercise on teachers working with students to interpret one or more texts the students had read). A 15- to 20-minute, continuous tape of such teacher/student interactions is generally quite revealing.

Viewing teachers' actions in this manner and joining such evidence with teachers' justifications for their actions and interpretations of student performance can prove especially telling. There will be questions about how much evidence is enough; about how much burden can be placed on teachers; and about the trade-offs in validity, reliability, and administrative feasibility between this proposal and some form of on-site observation. Both approaches deserve consideration, as does the question of paying teachers to assemble a mini-portfolio versus paying observers from outside their district to conduct on-site observations. Each approach has advantages and disadvantages. NBPTS has opted for the former because it allows the National Board to collect multiple samples of practice at low cost, avoids the idiosyncracies of schooling that can disrupt the class on any particular day, and creates a permanent record of teaching that can be referred to as many times as necessary if different judges take markedly different views of the practice or if, at a later time, there are legitimate grounds for an appeal.[2] However, the purposes of NCES and the National Board in collecting such data are significantly different and, consequently, replicating the NBPTS methodology may not necessarily be the best course.

Whatever method of observing practice is chosen, over the next several years the NBPTS system for advanced certification will produce a growing cadre of National Board Certified Teachers in multiple fields, some number of whom have been trained to score such evidence produced by peers and have also been found to be fair and reliable judges. The presence of this cadre provides NCES with a head start in being able to evaluate the data the Center could collect on teacher quality.

Moving forward in this direction would mark the first steps in developing a regular report to the nation on *The Condition of Teaching*. Knowing far more about the quality of teaching in particular fields (e.g., which aspects of science teaching teachers seem to have good command over and which they do not) would provide valuable guidance about where resources for teacher education need to be directed, not only within particular fields but also between fields as well. Understanding teachers' command of subject matter as well as the extent to which their beliefs about pedagogy conform to the current professional consensus in their field would be especially illuminating. Such data might move preservice programs to rethink some of their underlying assumptions, alert teachers and administrators to areas of potential weakness, and cause school districts to better target their scarce resources on professional development initiatives where the need is greatest.

The trap door to avoid in conducting such work is the trivializing of the complexities of teaching that can occur in a mad rush to design super efficient instruments. NBPTS has worked this territory hard to create professionally acceptable and administratively feasible assessments. On this score, the National Board has learned several lessons that should serve NCES well, including the need to avoid overly atomistic measures of practice that artificially disassemble the components of practice to the point where they lose their meaning. Recognizing that for many instructional issues there are multiple sound approaches, that judgments about the efficacy of teachers' actions need to take account of the instructional context, and that maintaining the authenticity of practice is especially important are other key considerations that must be attended to in such work.

Qualitative Measures of Institutions

However formidable these ideas for taking periodic measures of the quality of teaching may seem, they pale in comparison with designing a parallel plan for institutions and organizations that supply teacher education, which by virtue of performing this function play a significant role in shaping the kind and quality of teaching found in the schools. While attempting to develop such institutional measures may appear to be a fool's errand in light of the allergic reaction demonstrated by many institutions to any serious form of quality assurance (one need only have observed the withdrawal of colleges and universities from NCATE when faced with the reality or the potential of a negative finding), the uneven quality of teacher education that careful observers report they find nationwide suggests that progress on this front could have substantial benefits. The large number of institutions providing teacher education also argues for serious attention to this issue, as the proliferation of programs has raised well-founded concerns that weak programs soak up resources that might better be concentrated on upgrading those with greater potential for quality instruction; and in the bargain, the overall quality of education that is provided to prospective teachers is being compromised. The idea would not be to rank or rate individual institutions, but to provide a portrait with a good deal of fidelity that would accurately reflect the current state of teacher education.

What might such an effort be? At a minimum, it should consist of a set of careful case studies that NCES would regularly mount that would yield thoroughgoing portraits of the range of programmatic approaches being undertaken not only in preservice education but also in continuing education.[3] The latter arena is extraordinarily fractionated as post-licensing education takes place in traditional college and university settings, as well as in teacher centers, in school districts, in seminars and courses run by disciplinary and specialty groups, and in other informal settings. But this diversity of approaches should not discourage NCES, because it is in just such settings that some of the best and some of the worst teacher education is located. In addition, it is in the quality of such post-licensing education that the prospects for a novice teacher to 1–day advance to expert status are forged. So it is an arena that is crucial to the health of the profession and deserves much more attention than it is has traditionally been accorded.

NCES should also consider moving beyond the case study option to a format that more closely parallels the studies of teaching quality suggested above. While this approach would represent a sharp departure from current practice, it should not be dismissed out-of-hand because its potential to illuminate this critical aspect of teacher education is substantial.

What better way to judge the quality of teacher education and disarm those critics who claim it leads nowhere than to have NCES engage an independent auditor with some stature to visit a good cross-section of the nation's colleges of education and look beyond the plans, the curricular offerings, the admissions and graduation requirements, library holdings and faculty credentials to the actual teaching of prospective teachers that takes place? The objective would be to paint a picture of the kind and quality of preparation and ongoing professional development that the nation's current and prospective teachers are experiencing whether in schools of education or in colleges of arts and sciences.

To conduct such work, standards of teaching teachers would have to be developed, and this task should not be conducted by NCES but by some independent entity with fiscal support from NCES.[4] This exercise in and of itself would be healthy for the entire higher education community, as it would establish a set of commonly shared principles that would frame a host of instructional and curricular decisions that college and university faculty have to make on a regular basis. It would also provide guidance for designers and sponsors of continuing professional education and for consumers of their services.

Establishing standards for teaching teachers would also be especially timely given the efforts of several reform initiatives to provide the enterprise with substantial uplift by placing greater emphasis on clinical training, on the closer integration of subject matter and pedagogy, on developing the habit of self-examination, and on the translation of theory into practice through professional development schools (i.e., the educational equivalent of teaching hospitals) and other vehicles.

Flows Through the System

As student enrollments have ebbed and flowed, the demographic turmoil that was periodically provoked in the schools led inexorably to a growing interest in data on teacher supply and demand. However, in an environment where teacher quality is not an issue, the supply of teachers is inexhaustible. Still, the devotion of NCES to collecting data on the movement of teachers within the system has been rewarded with findings about how minimal standards can be manipulated in the face of potential teacher shortages—an underreported scandal in American education. Whether it is the issuance of emergency licenses, the lowering of entry-level standards of competence, or the misassignment of teachers (an example of a low-cost proxy for the larger qualitative problems that plague teaching), such threats to the quality of education America's children receive need to be closely monitored.

Improvements in the ability of NCES to understand the complexity of teacher flows both within and in and out of the teacher work force should now permit finer grain studies of change over time and the effects of changes in policy on such flows. Of special interest on this score is the degree to which the professionalization of teaching may affect who is attracted to teaching, who stays, and who leaves. Today, the incentive structures in most local systems encourage teachers to become administrators, resulting in the loss of some of the strongest practitioners from the classroom. National Board Certification and other worthwhile initiatives are making a concerted effort to change this reality and to change the culture of teaching to provide better support to novice and experienced teachers, to value continuous learning and problem solving in the company of one's peers, to recognize and reward excellence, and to alter schools' organizational structures to capitalize on the knowledge, skill, and expertise of the profession's strongest practitioners. If these efforts begin to take hold, teaching should become a much more attractive career option, and its ability to attract and hold talented people with many other career choices should improve.

To the extent these changes materialize on a large scale, they will do so gradually, and they will most likely occur unevenly across the education landscape, as risk taking of this sort is not the common condition usually found in the schools. Thus, it will be important not just to be able to track the aggregate flows of teachers in and around the system, but also to look for changes in patterns linked to changes in policies of the sort noted above. In the short term, this may mean more targeted studies focused on those states and localities where these initiatives are beginning to take off.

In the various efforts that NCES may undertake to track teachers, special attention should be given to the dynamics of minority teacher movements, because there remains a sharp disjuncture between the percentage of the teaching force that is drawn from the minority community and the percentage of minority students in the schools and the percentage of people of color in the labor force. This issue needs attention not because minority children must be taught by minority teachers, but because the teaching force itself would be strengthened by teachers with more diverse life experiences and all students would benefit from working with a range of adults who are more reflective of the larger society they are being prepared to enter. It is important to understand how the career paths of teachers of color and those of their white counterparts are similar or different, and to understand to what extent these differences are either being magnified or dampened over time.

All such work will be markedly enhanced as the numbers of National Board Certified Teachers (NBCTs) begin to multiply, because it will allow analysts to bring a previously unattainable qualitative dimension to their studies. As NCES thinks about future requirements, planning to track the movements of NBCTs should be a high priority. The kind of education they have had throughout their careers, the nature of their teaching experiences, the conditions of their workplaces, and the changes in their employment circumstances over time should prove especially illuminating. By following this course, clues to the merits of various teacher recruitment, hiring, assignment, and compensation policies would be subject to more searching investigations, and the question of the uneven distribution of teaching talent could be given the concerted attention it is due. On this latter point, it is well understood by those close to the system that inequities in school finance and other factors result in some schools having their pick of exemplary teachers whenever a vacancy occurs, while others serve as farm teams for their more advantaged colleagues. Lacking valid measures of teacher quality, such discrepancies have been too easy to overlook. With the advent of National Board Certification not only can sound analyses be conducted on this point, but also the effects of new policies to ameliorate this problem can be reasonably judged.

In time, one would also want to examine the effect of NBCTs on the practice of those who work in close proximity to them, but such studies probably are best sponsored by the research arm of OERI.

Kind and Level of Investment in Teacher Quality

Although the nation is fortunate to have many excellent teachers in its schools, there is no question that it needs more and that the overall quality of teaching needs to improve significantly. This will not happen overnight, nor will it happen without a concerted effort by many parties. In fact, it will require a sea change in labor-management relations, a break from past practices along the entire front of teacher education, and a willingness by states and localities to commit to a dramatic reconception of how to foster the growth of exemplary practitioners. This means that schools will have to find better ways to invest the resources they currently allocate to teacher education (not treating it as a fringe benefit and not just paying for the accumulation of graduate credits at the cheapest and most convenient institution teachers can find), while simultaneously increasing their investment in their most valuable capital assets, teachers.

Just how school districts take up this challenge is crucial, and the extent to which teachers meet their employers halfway is no less important. NCES can contribute to this effort by monitoring it closely through the Schools and Staffing Survey or through special supplements to it. Some of the more interesting questions on the table where data collection and analyses could be especially telling are the following:

- What are teachers doing to strengthen their practice? How much effort are they expending in this direction? How much support/encouragement are they receiving in this direction, and what form is it taking?

- What is the level of public sector investment in initial teacher preparation and ongoing professional development? Where are these investments being made, and what are they purchasing?
- What is the level of public and private sector investment in research and development to improve teaching?

Pursuing these questions and others that they generate should provide NCES with a rich but manageable challenge. Unlike the dicier questions of measuring teacher quality discussed above, these matters should yield to the more common methodologies at which NCES is well practiced.

EPILOGUE

Education has been through a period where it has been subject to high-level, high-intensity, and high-volume scrutiny. While it has been painful for some, it is an extraordinarily promising exercise for the country because it is a sign that the nation understands that excellence and equity in education are necessary prerequisites to a healthy society. As this debate has proceeded, there appears to be a growing recognition of the centrality of the quality of teaching and of the ways in which the education and training of teachers proceeds. Surprising as it may seem today, this was not the focus of many early efforts at education reform in the 1980s, and it remains overlooked by some to their detriment even in the 1990s. Still, as this paper is written, we are just beginning to get our feet wet in this crucial arena of teacher quality.

With the advent of more valid measures of teacher knowledge and performance now emerging from the profession, NCES is well positioned to harvest the work of those toiling in the teacher quality vineyard and to begin new streams of data collection that will take advantage of these path-breaking efforts. This is work that a federal agency could not have conducted on its own without encountering substantial legitimate opposition. But now that others have taken the lead in defining the parameters of highly accomplished teaching, there exists a special opportunity to advance the work of NCES that should not be missed. And, that is to join in pushing the frontiers of this field of education measurement forward by investing in valid, reliable, and efficient means of gathering information on the quality of teaching and on the quality of the system support structures that can contribute to developing the excellent teachers America's schools desperately need in much greater numbers.

One should acknowledge that much of what is proposed in this paper is a significant departure from standard operating procedure at NCES, but the unavoidable fact is that these procedures in many respects are extraordinarily limiting. If NCES is only to continue along the current track it is following, it will be pursuing a strategy that will not yield much of the data the country requires to address some of its most difficult education problems. However, expanding its vision in some of the ways suggested here, or in other ways but with the same objectives in mind, should increase the prospects that key public policy decisions will be grounded in hard won intelligence about the system's characteristics and be less subject to influence by ignorance, guesswork, or untested theoretical constructs. If such a result ensues, American schools and students will be the beneficiaries.

NOTES

1. For example, only a few states have begun to insist on observing novice teachers during their early years of practice as part of their determination of competence before awarding a full-fledged license. Thus, being able to assure the public that teachers are able to convert theory and knowledge into sound practice in the early years of their practice is a warrant that accompanies only a modest fraction of the state licenses in circulation today.

2. Teachers also find this process commendable on several grounds. They get to put their best foot forward. They find it a means to open a new professional conversation with colleagues about the trials, tribulations, and achievements of their practice. And, they find the process fair, in part, because all of the judgments are made by peers from outside their school district.

3. This proposal would clearly be a departure for NCES, as such work has historically been sponsored by other OERI entities rather than the one charged with collecting, analyzing, and interpreting statistical data.

4. This is a model similar to that invoked to provide federal support for the discipline-based student standards and standards for advanced certification of elementary and secondary school teachers.

REFERENCES

National Board for Professional Teaching Standards. 1994. *Early Adolescence/English Language Arts Standards for National Board Certification*. Detroit, MI.

National Board for Professional Teaching Standards. 1994. *Early Adolescence/Generalist Standards for National Board Certification*. Detroit, MI.

National Board for Professional Teaching Standards. 1995. *Early Childhood/Generalist Standards for National Board Certification*. Detroit, MI.

National Board for Professional Teaching Standards. 1996. *Middle Childhood/Generalist Standards for National Board Certification*. Detroit, MI.

National Board for Professional Teaching Standards. 1991. Chapter II: *What Teachers Should Know and Be Able to Do. Toward High and Rigorous Standards for the Teaching Profession: Initial Policies and Perspectives of the National Board for Professional Teaching Standards*. Third Edition. Detroit, MI.

Discussant Comments

MICHAEL TIMPANE

We are at a time when education policy presumes to say that it will change students' achievement outcomes measurably, either through systemic policies that clearly link standards, curricula, school programs, pedagogy, and assessments, or through the interplay of market forces in various systems of school choices. It is also a time when researchers are painstakingly establishing the combinations of content, pedagogy, and learning that characterize particularly successful classrooms and schools. In these circumstances, it is of the greatest importance that policymakers and program planners have reliable data on patterns of classroom activity and teacher performance, to know where we started, how things change, and with what effect. Thus, NCES should accord a very high priority to pursuing the lines of development suggested by Brewer and Stasz and by Mandel.

That said, the authors, correctly ambitious in laying out the possibilities, are equally correctly ambivalent about the prospects for success. They are correct to wonder how much of the needed data could be collected by survey instruments, however subtle and sophisticated they may be; the examples they present from recent studies are as halting as they are hopeful in sketching the pace of progress. All in all, their papers raise a formidable list of conceptual and empirical concerns:

- How can such data collections reliably connect teaching practices with levels or patterns of learning? And how should such learning be defined and measured?
- To what extent must such surveys be preceded and/or complemented by other small-scale studies involving interviews with teachers, classroom observations, review of logs, artifacts performance assessments, and so on?
- How would data collection and analysis take account of explanatory policy variables, such as teacher preparation and rewards, incentives and capacities for change, legislative and regulatory constraints ranging from teacher certification to union contracts to federal and state program requirements?

Staggering though the designs problems may be, let the predictably lengthy development process begin. We must have more compelling information about the ways in which our expectations—as set forth in polices and paradigms of teaching and learning, backed up with resources—are realized in schools and classrooms, and with what effect. Otherwise, our research and data collection will seem increasingly marginal to educational policy and practice, and educational policy and practice may seem that much more marginal to our nation's progress.

EILEEN M. SCLAN

Both the Brewer and Stasz and the Mandel papers suggest that NCES look at the multidimensional nature of school processes and how interactional effects of the teaching and learning process either enhance or reduce opportunities to learn. I support this effort. In this response, I will focus on two main issues: 1) the importance of contextual data and 2) its relevance to equity issues. Collecting contextual information about curriculum, student learning processes, instructional resources, and professional development of teachers requires a new way of thinking about data collection in schools. The challenge for NCES is to move beyond gathering data that reflect a static view of teachers' and students' realities toward collecting data that captures a dynamic view and that accounts for contextual variation and the vital nature of the teaching and learning processes. To begin the unprecedented task of gathering rich, detailed data on what goes on in schools and classrooms will require using non-survey methodologies.

Because teaching requires active involvement and is not something that is done to students, data are needed on student characteristics, the teaching and learning process, and teacher-student interactions to provide useful information to researchers, policymakers, and teachers themselves in understanding teaching and learning in its richer context. What students think, what they feel, and what they experience in and out of school influence the way they learn and how well they learn. In response to this need, innovative teachers have begun to implement more authentic assessment measures to better understand their students' learning. Both papers discuss the importance of employing authentic methods of assessing student and teacher performance through the following methods: teachers' work samples; logs; teachers' judgments in individual situations; the context of the teaching situation; interviews; observations; videotaping; and artifacts such as homework assignments, quizzes, texts, and exams. Qualitative methodologies may provide a closer look at how teachers can nurture students in becoming active, responsible, and responsive decisionmakers.

Workplace structures also influence the prevalence and nature of teacher and student opportunities to learn. In Rosenholtz's (1989) seminal work on teachers' workplaces, for example, the extent of teachers' learning opportunities was associated with increased student achievement gains. The nature of the social structures, policies, and traditions of the school environment often determine teachers' opportunities to learn. Teachers who have opportunities themselves to stay abreast of the exploding knowledge base in developmental psychology, social organizational theory, state-of-the-art teaching techniques, and subject matter content are more likely to provide opportunities for students that enable them to think critically, creatively, and deeply. Mandel aptly points out that investments in ongoing teacher education are key determinants of the quality of teaching. Teachers say they need institutional supports, such as time to meet with colleagues and to participate in policymaking at the school or district level, access to instructional materials and to recent research journals, and opportunities to attend professional conferences. To understand schools, we must understand them as teachers and students do. The authors call for a greater focus on the dynamic social organizational dimensions of workplace incentives that support teachers in becoming more effective, reflective, and analytical practitioners.

Contextual data on the wide spectrum of student experiences in classrooms, the nature of their opportunities to learn, and how they process particular forms of knowledge delivered in varying ways will provide us with windows to view the complexities of teaching and learning.

This information will allow us to study the more subtle interactional effects between teachers and students and workplace environmental variables, and how these factors, in turn, contribute to teaching for higher level problem solving, analyzing, critical and creative thinking, and deeper levels of appreciation for subject matter, issues, and people.

Focusing on contextual variables is even more important when we consider equity issues. Teachers are expected to deal with complex challenges in that children bring their social lives at home, on the street, in their communities, as well as their feelings about themselves as learners, to the classroom. Opportunities for teachers to learn becomes a critical issue at a time when increasing numbers of students are walking into classrooms who have been neglected, abused, or deprived, and whose experience with adults and schools set up and reinforce a sense of futility, which comes out of a long history of expectations of failure. Brewer, Stasz, and Mandel speak to the need to gauge excellence in teaching and to track the equitable distribution of opportunities to learn for teachers and students. The most daunting problem that we as educators face heading into the next century is one of achieving equity—students' equal access to knowledge and learning experiences. Gross inequities of resources have been documented in case studies by Kozol and in national data in the mathematics and science teaching fields by Oakes. More recently, my work with Darling-Hammond validates what we already know through anecdotal evidence—that is, students in poor and minority schools are taught by the least qualified and most inexperienced teachers.

Brewer and Stasz underscore the importance of ensuring equitable delivery of schooling for all students. Mandel calls attention to the underreported scandal in American education: in the face of shortages of qualified teachers, we are opting for short-term solutions by issuing emergency certificates, by lowering entry-level standards, and by misassigning teachers out of their fields. If disproportionate numbers of the least prepared elementary and secondary public school teachers teach in the most disadvantaged communities, it is likely that these children are experiencing fewer opportunities to learn than children in the most advantaged communities.

Thus, data that illustrate the interactions between teacher performance, student learning processes, and workplace supports may help us to deepen our understanding of the complexities of teaching and learning and also to document equitable delivery of schooling to all students.

References

Darling–Hammond, L. and Sclan, E.M. 1996. "Who Teachers Are and What They Think." In T. Butterly and J. Sikula (eds.), *Handbook of Teacher Education.* Association of Teacher Educators.

Kozol, J. 1992. *Savage Inequalities.* New York: Crown.

Oakes, J. 1990. *Multiplying Inequalities: The Unequal Distribution of Mathematics and Science Opportunities.* Santa Monica, CA: Rand.

Rosenholtz, S. 1991. *Teachers' Workplace.* New York: Teachers College Press.

MARY ROLLEFSON

David Mandel begins his paper with a "chicken or the egg" problem: Without a definition of the "elements of highly accomplished practice" we do not know how to collect data on teacher and teaching quality, but without data on teaching and student learning we can not define teacher quality. In other words without a definition, we can not collect data, and without data we can not validate a definition.

Responsibility for this situation is put on the teaching profession, but I believe that responsibility extends to the education research community as well, which has found little empirical support for the relationship between teacher and teaching variables and student outcomes. However, as Emerson Elliott mentioned at the conference, research from the University of Wisconsin Center on Authentic Pedagogy, which identifies qualities of teaching that relate to improvements in student performance, holds promise.

Mandel proposes the standards of the National Board for Professional Teaching Standards as the framework for better data on the quality of teachers and the institutions and the programs designed to provide them. In addition, he emphasizes the importance of data on teacher quality and suggests that all of our data on teachers are marginalized and diluted, without quality information that goes beyond our standard criteria of major or minor and certification in the teaching assignment field.

It is true, as Mandel says, that NCES "stubs it toe" on this issue and that quality data would vastly improve our understanding of teacher supply and demand—which ultimately is a quality issue, since imbalances are often resolved through adjustments in teacher quality. Another area where teacher quality data would improve understanding is in equity. That is, among different populations of students, which population has access to the type of quality teaching that makes a difference in learning? I suspect that important differences in the quality of teachers from different supply sources and in student access to good teaching are only hinted at in our current data on teachers.

Good measures of teacher quality would also help us understand how quality practice develops or fails to do so both in the course of teacher education and throughout a teaching career. The implications of this understanding for teacher education and continuing professional development are enormous.

Mandel also points out the importance of quality data for examining institutions of teacher education. In terms of a future direction for NCES, this may be important as well. I would suggest looking to NCATE standards and the work of INTASC (Interstate New Teacher Assessment and Support Consortium) for a framework. But if teacher quality is in its infancy, I suspect much work needs to be done in this area.

His recommendations to NCES are twofold: one is the use of the more traditional method of a statistical agency, the other relies on methods that are less traditional—case studies and assessments of teacher preparation programs. The latter calls into question what our appropriate role is, especially in a time of budget cuts. However, these approaches are more appropriately the territory of OERI and the wider education research community. In addition, Mandel points out the critical connection between research and statistical data collection. I think NCES would eagerly await these results.

I believe that some of the more traditional recommendations hold promise. For instance, one is to collect teacher data every 5 years, which we already are doing in SASS, although the periodicity varies. Related to this is the recommendation to combine National Board information on board-certified teachers—rich information from teacher videos and commentaries on their teaching goals, context of instruction, student understanding, and the teacher rationale for the approach they use—with NCES data collection on these teachers, and to follow these teachers. Also, current plans have been made to include board-certified teachers in the next SASS, and perhaps they could also be followed in the Teacher Follow-up Survey. The combination of a sample of *quality* teachers in a data set that contains both rich National Board data and standard SASS data would provide us with an opportunity to do research on the issue of teacher quality. Finally, another good recommendation is to use board-certified teachers to help NCES assess its data on teacher quality.

Thus, as mentioned previously, without defining teacher quality we can not collect data and without data we can not validate the definition of teacher quality.[1] These two conditions define the ground for basic research; using that research as a framework, indicators of teacher quality can be developed to collect data from large nationally representative samples. This needs to happen before we can apply our usual methodologies to this resolving issue.

Mandel's paper raised several questions, which I would like to comment on. First, if National Board standards are a framework for defining teacher quality, then what are the dimensions of that framework, and what are some of its most important content areas? I would like to have had those laid out more explicitly in his paper. If we are not yet close to a working definition, then perhaps we can at least get a hint of where we are now and how far we have to go. I would also suggest adding some other places to search for definitions of teacher quality: for example, the NCATE standards, the work of INTASC (the Interstate New Teacher Assessment and Support Consortium), and some of the OTL research discussed in the Brewer and Stasz paper. Furthermore, we need more qualification about teacher quality—that is, we need data not just on the excellent, highly accomplished, and advanced end of the continuum, but on the whole distribution.

Finally, there were some more practical recommendations that were not discussed as thoroughly. These were to collect data on:

- Teachers' professional development—What they are doing to strengthen practice, and what kind of support are they receiving to do it?
- Public investment in teacher preparation and in continuing professional development—What kinds of investments and what kinds of programs? This points to the need for basic data on institutions that educate teachers.

Although these two items are provided almost as an afterthought, they point to important areas in which data could be improved in the near future. And they certainly deserve more discussion.

Notes

1. When developing a working definition of teacher quality, it would be useful to turn to the standards of the National Board for Professional Teaching Standards.

SHARON BOBBITT

On the first day of school in the fall, both new and experienced teachers go into their classrooms to meet their new students. They shut the door, take roll, and begin practicing what has been called "the second most private act"—teaching. While this stereotype is slowly changing, little is known about what happens inside the classroom walls, in the interaction of teacher and student, to bring about learning. The Brewer and Stasz paper could have been entitled "Illuminating the Black Box." The paper makes a serious and credible attempt to help the National Center for Education Statistics figure out how to measure and report on what happens inside the black box of the elementary/secondary school classroom.

Brewer and Stasz use the terminology of Opportunity to Learn (OTL) to discuss data issues related to instructional practices and classroom processes. OTL started as a narrow concept in the context of international surveys. In reporting achievement scores in mathematics across many countries, a key variable was whether students in each country had been exposed to mathematical concepts in their curriculum by the time the assessment was conducted. Obviously, a student who has never had the opportunity to learn a mathematical concept will perform less well on those portions of the assessment that tap this concept than a student who has had extensive exposure to the concept in his or her school curriculum. The SIMS study asked teachers to rate whether or not their students had been exposed to the items in the assessment. This narrow definition of opportunity to learn was simply, therefore, a measure of students' exposure to items on which they were being assessed.

Andy Porter and others have taken the concept of opportunity to learn and expanded it to encompass a much broader definition. Porter argues that a student's opportunity to learn includes not only appropriate curriculum content but also high-quality pedagogy and adequate classroom resources. Students who have access to more of these three elements have, in some ways, more opportunity to learn challenging content. It is this framework of OTL that Brewer and Stasz adopt to discuss the measurement of classroom processes by NCES.

Using this OTL framework, Brewer and Stasz make four recommendations. First, they suggest that NCES enhance survey items on curriculum content and pedagogy. Second, they recommend enhancing our measurement of student learning through survey and collecting artifacts. Third, they suggest enriching classroom process data through alternative data collection methods. Finally, they suggest enhancing instructional resource measures. I would like to address each of these recommendations.

One of key issues abut measuring curriculum content and pedagogy, especially in nationally representative surveys, is whether to make the questions subject- and grade-specific or try to formulate items that would be applicable to all teachers. Policy Study Associates (PSA), under contract to NCES, has recently struggled with this issue for the last administration of the

Teacher Follow-up Survey in school year 1994–95. This questionnaire is administered to a nationally representative sample of teachers of all subjects in grades K through 12. NCES wanted to include a module of questions on teacher instructional practices and classroom processes that would be applicable to the diverse group of teachers in the sample. Building on previous research in this area by Porter Burstein and McDonnell and others, PSA developed a set of questionnaire items that were tested in teacher focus groups before being administered in the TFS. It will be very interesting to see the amount of variation in a sample of over 7,000 teachers in basic instructional practices. The data will be available in the spring of 1996. While instructional practices may be able to be generalized to apply to all teachers, curriculum content must of necessity be subject- and grade-specific. NCES has also recently funded some additional work by Andy Porter on his teacher questionnaires about curriculum content for middle school math and science. By attempting to develop survey items that will be broadly applicable where possible and subject-specific when necessary, NCES hopes to improve the measurement of curriculum content and instructional practices in its large-scale surveys.

The second recommendation involves enhancing measurement of student learning through surveys and the collection of artifacts. As the authors correctly note, you cannot measure everything you want to know through large-scale surveys, Although it is very early in the process, NCES is experimenting with other forms of data collection to enhance our understanding of things that surveys cannot measure. The authors point to the work of Jim Stigler, who has videotaped teachers in their classrooms, as a promising possibility to build our capacity to measure what is going on inside the black box. It is also possible for NCES to fund, support, or conduct smaller scale research efforts that feed off of the large-scale sample surveys. Targeted subsamples (or even nationally representative subsamples) could be selected from existing sample surveys, such as the Schools and Staffing Survey, to investigate issues that are either too complex or too expensive (or both) to measure on the entire sample. Such subsamples could be candidates for using innovative data collection methodologies such as videotape (like Stigler), teacher logs, or classroom observation.

Thirdly, Brewer and Stasz recommend enriching classroom process data through alternative data collection methodologies. Working again with PSA, NCES has funded the development of survey items intended to measure instructional practices and classroom process in middle school mathematics. While the exact content of these items is not important, the process of development is. PSA is using teacher logs, classroom observation, and focus groups to validate the items that the teachers fill out in the survey form. The development of a validated module of items would enhance the ability of NCES to measure instructional practices in all of its large-scale surveys.

Finally, the paper recommends that we enhance our measures of instructional resources. I agree that this is one area where our surveys are weak, yet we have not made too much progress to date in working to fill this gap. One of the most outstanding teachers in the country recently told me at the Goals 1000 Teacher Forum that she gets $40 each year from her school for classroom supplies. Teachers on the U.S. Department of Education listserv report routinely spending $1,000 to $2,000 of their own money each year for instructional supplies. Brewer and Stasz coin a wonderful term for this phenomenon, the "opportunity to teach." NCES can add and should improve the measurement of the resources available for instruction in this country's classrooms.

The Brewer and Stasz paper leaves us with several questions about the relationship of instructional practices and student outcomes. How do you validate measures of classroom processes and their impact on student learning, broadly speaking? Is it enough that teachers are doing the things in the classroom that the experts (for example, NCTM and NSTA) say they should be doing? Or does there need to be evidence that the pedagogy is resulting in improved student outcomes? Are instructional practices, in and of themselves, worth studying? Hopefully, as we improve our ability to measure classroom processes, we will be able to understand better the complex relationship of curriculum content, instructional pedagogy, and resources, and we will be able to get a better glimpse inside the black box of the classroom.

4

Trends in Statistical and Analytic Methodology: Implications for National Surveys

"So What?" The Implications of New Analytic Methods for Designing NCES Surveys

Robert F. Boruch
George Terhanian

SUMMARY

This report was commissioned to address the question "How can advances in statistical analysis be used to improve the design of surveys?" The surveys of paramount interest are those sponsored by the National Center for Education Statistics (NCES). The "advances," as initially conceived, include new approaches to analysis that have been invented by statisticians, mathematicians, and methodologists.

Advance: Mathematical statisticians and methodologists, at times, remarkably improve the way we analyze statistical data. But they rarely describe how their advances can improve the *design* of surveys. Scholars who apply the new (or old) methods to NCES data, at times, speculate on how NCES surveys might be improved and report their suggestions in journal articles.

Implications: First, NCES can encourage scholars who invent new analytic approaches to educe the implications of their advances for improving survey design. NCES should not expect to find explicit implications absent such encouragement. Second, NCES can encourage scholars who apply new (or old) analytic approaches to NCES data to educe the implications of their results for better survey design and to contribute more effectively to a common pool of implications. Third, NCES can exploit mechanisms that NCES and other federal agencies already depend on to build this knowledge pool, e.g., external committees and internal staff. Fourth, NCES may exploit new technology to do so, notably on the World Wide Web (see section on New Technology).

Cross-Design Synthesis

Advance: Recent work on cross-design synthesis suggests that, at times, survey-based studies of the effect of national programs, based on probability samples or administrative records, can be combined with local controlled experiments on the programs' effects so as to produce better national estimates of the impact of the programs. In the long run, combining such information is arguably important to advancing knowledge and to the efficient exploitation of resources in both the survey sector and the experimentation/evaluation sector.

Implications: First, to foster good cross-design synthesis, NCES surveys can be designed so as to permit linkage of the surveys to controlled experiments. Experiments at the local level, over which

NCES has no direct control, can be designed so as to permit linkage with NCES surveys. That is, both the surveys and the independently conducted experiments can be designed cooperatively so that response variables, treatment variables, target populations, and propensity variables are measured in the same way. Second, NCES can ask, or learn how to better ask, about propensity so as to enhance analyses and synthesis. NCES can do so in ways that others have not, through cognitive research and other approaches.

Hierarchical Models, Models in General, Theory, and the Design of NCES Surveys

Advance: Hierarchical models and associated models and analysis help to frame the way we look at data that are generated at the national level, state level within nation, school district level within state, classroom within school or district, and children within schools, and the way we examine data on each child or classroom, and so on across a time frame. A notable advance lies in contemporary software.

Implications: The claims made by the developers of hierarchical models are sufficiently broad as to allow vague statements about how new NCES or any other surveys should be designed. A first such implication is that NCES should collect multi-level data, as it has in the past. A second equally vague implication is that the NCES effort ought to be expanded, invigorated, and made more disciplined in the context of HM technology, e.g., figuring how whether and how to enlarge sample size at certain levels. Although proponents of HM may merely identify general implications, at least some who employ the approach are more specific. A third set of implications is that NAEP 1) should measure socioeconomic status more directly or less indirectly than it now does; 2) should get at teacher instruction variables better; 3) should elicit information from more teachers if indeed we want to know about their influence; and 4) may have to sample more students within schools. A fourth design implication for NCES is that investments have to be made in understanding how to estimate sample size within each level in a hierarchical scenario. A fifth implication is that NCES has to decide where HM-driven implications ought to be exploited, e.g., in designing NAEP versus NELS:88. Other simpler models and analytic approaches may be better and, in any case, theory ought to drive some of this. NCES has to take theory into account somehow.

Advance: Meta-analysis, which can be construed in terms of hierarchical models, involves the combination of multiple studies.

Implications: NCES surveys can be designed so as to exploit the results of meta-analyses to design a survey. This requires, in the design or modification of each survey, the invention of a mechanism for linking the survey at hand to other related surveys or experiments (Sections 2 and 6).

Counting the Hard to Count, Measuring the Hard to Measure

Advance: New developments in analyzing count data suggest that a social network-based estimator of the incidence or rate of a sensitive behavior can, at times, be informative. Such estimators avoid certain privacy problems in educational and social surveys, and avoid the appearance of problems. That is, they are based on questions asked about unidentified people, not on questions about the respondent's own potentially embarrassing behavior.

Implications: When privacy is an issue but understanding the incidence of a sensitive behavior is important, NCES can consider the design surveys that exploit network-based estimators. A second implication is that some basic research, pilot work, and verification research are, as usual, necessary.

Advance: Cognitive approaches appear not to have been employed often in test development despite their use in other survey areas. Full information matrix factor analysis is alleged to be a relatively new way to get at the structure underlying test results. Neither analytic approach itself has obvious implications for NCES survey design. A Stanford group used these, together with other methods, and applied them to mathematics and science data and other information from NELS:88. They produced implications for design of NELS:88 and perhaps other surveys.

Implications: Mathematics reasoning and knowledge are two distinct latent factors underlying test scores generated in NELS:88. They ought to be treated as such inasmuch as total scores are arguably misleading. Science scores are characterized by many different factors. Moreover, each type of factor is influenced in theory and predicted empirically by different variables whose measurement in NELS:88 can be improved. Some variables that may relate differently to each factor are not measured at all, e.g., instructional practices such as discovery learning or reciprocal teaching. A main implication is that NCES can exploit theory of how knowledge and reasoning are affected by various factors. The theory and analyses can be used to drive NCES decisions about what to measure, how deeply to measure, and why.

Small Area Estimators, and So On

Advance: Recent work on indirect estimators suggests that it is possible, at times, to develop good small area estimators based on 1) data from a national probability sample, 2) information obtained independent of the national sample, and 3) a model that links the two. "Good" estimators here means that they are more plausible than any alternatives.

Implications: NCES surveys can be designed so as to exploit new work in domain indirect, time indirect, or time and domain indirect estimators. Time indirect estimators might be tested to understand whether they suffice to permit reducing NCES annual data collection efforts to biennial efforts, or to lengthening the time between points of measurement in NAEP and other periodic surveys. Domain indirect or time indirect estimators might *now* be tested to determine if satisfactory local area estimators can be produced or if certain area surveys now producing direct estimators can be reduced. Validation tests are possible because NCES now relies heavily on direct estimates.

Satellite Policy

Advance: NCES survey data are at times used to sustain analyses of cause and effect. The problems in doing so are complex, numerous, and have been discussed often and in numbing detail.

Implications: NCES surveys can, at times, be designed to facilitate local controlled experiments, for example, by oversampling the subgroups that are targeted for experimental programs. This requires survey designs that permit linkage between the surveys and experiments.

Linking NCES Surveys and Data Sets from Other Sources

Advance: Multiple independent surveys are undertaken often, and with good reason, by NCES and various other federal agencies. To judge from recent analytic work, the independence of surveys mounted by different agencies or units within the Department means, however, that the results of different surveys often cannot be easily integrated, compared, combined, or otherwise linked. More important, NCES has had substantial recent experience in the problem of integrating certain data collection efforts, e.g., the CCD.

Implications: NCES can take a leadership role in learning how to run independent surveys or studies more generally so that linkage, comparison, integration, or merger is possible despite their independence. The task hinges on enhancing the extent to which major factors are common to different databases, e.g., variables, ways of measuring the variable, target population. It hinges on the invention of ways to specify the lack of commonness, and on the invention of ways to induce artificial commonness.

New Technology

Advance: The development of the Internet, especially the World Wide Web, does not fall into the category of advances that concern us here. Nevertheless, it is too important to ignore.

Implications: There are a variety of tactics that might be exploited in interest of better design of NCES surveys. They include Web-based surveys of data users and analysts to 1) elicit direct information on questions, design characteristics, and so on; 2) build a registry of users, uses, and products; 3) distribute spreadsheet files; 4) track the emergence and development of new analytic methods; 5) create electronic discussion groups among analysts and designers; 6) post frequently proposed questions and their answers; and 7) exploit Adobe functions to better disseminate information.

INTRODUCTION

This report focuses on what new analytic methods imply for the design of better surveys. The surveys of special interest here are those conducted by the National Center for Education Statistics (NCES) (Davis and Sonnenberg 1993; Davis and Sonnenberg 1995).

The report's tôpic was determined jointly by the author, NCES, and an NCES contractor, MPR Associates. It was chosen to assure that NCES could exploit new opportunities to enhance survey design on education in the United States if indeed such opportunities are engendered by new analysis methods (NCES 1995). The various parts of this report vary in their length, developmental stage, and depth. Some are better thought out than others; some implications are stronger than others.

The first section examines the broad question: "What are the implications of new analytic methods for the design of NCES surveys?" It describes why the answers to the question are hard to produce. It also describes why and how implications can be produced.

The next section concerns recent work on cross-design synthesis. It argues that data generated in surveys of the sort undertaken by NCES can be combined at times with controlled experiments sponsored by other federal agencies or by private foundations. This combination of data is in the interest of better estimating the effects of federally sponsored education programs and policies.

The third section focuses on recent work on hierarchical models and other statistical models. Some implications are obvious, provided there is some agreement that measuring individual growth trajectories, or estimating the effects of schools is important.

Counting the hard to count and measuring the hard to measure is considered in the fourth section. We focus on network-based estimators and on recent analyses of NELS:88. NCES cannot always elicit information directly about the private behaviors of students, teachers, parents, and so on. This is despite the fact that these behaviors, such as criminal or sexual or disruptive activity, may be important on policy grounds. One new method, invented by a quantitative anthropologist and a physicist, is reviewed here and the implications are laid out. A second section covers the product of an interesting effort by Stanford scholars to learn how to improve NELS:88.

The fifth section of the report concerns indirect estimators, including small area estimators. The object is to understand how NCES, whose efforts are routinely based on large scale periodic national samples, can estimate the incidence of problems in small geographic areas or can abstain from one cycle of a national data collection effort. Achieving either object is not trivial, given NCES's mission to produce data based on national probability surveys and the pressure to say something at the subnational (small area) level, and given the pressure to produce information on a regular cycle, and given restricted resources.

Section six is entitled satellite policy. It argues that the NCES surveys and others ought to be an unobtrusive platform for controlled experiments run by other technical agencies or private foundations.

Section seven concerns the idea of linking surveys and data sets. Linkage, combination, comparison, and related ideas are considered briefly. This essay exploits research that was sponsored by the National Science Foundation and is relevant to NCES interests.

The last section of this report considers new technologies and how they might be exploited to enhance the design of NCES surveys. The focus is on the Internet and how the "Net" can be exploited in the interest of designing better NCES surveys.

EDUCING THE IMPLICATIONS OF NEW ANALYTIC METHODS FOR THE DESIGN OF SURVEYS: SOME PECULIAR DIFFICULTIES

The question at hand is "What are the implications of new approaches to statistical analysis for the design of surveys?" Put another way: "How can surveys be improved, based on advances in analytic methods?" A basic reason for posing the question is that it seems important. Or at least interesting.

The presumption is that an agency, such as NCES, can exploit advances made by the inventors of new ways to analyze data. A further presumption is that exploitation can enhance the design of the National Assessment of Education Progress and other surveys. It seems then sensible for the agency to do so. In the abstract at least, one might speculate that advances in analytic methods might for example, lead to designs that enhance the precision, informativeness, or usefulness of surveys or decrease their costs or difficulty. The phrase "in the abstract" is of course important here.

A second reason for asking the question has to do with an early partial flop. A decade ago, a Social Science Research Council Committee on Evaluating Longitudinal Surveys addressed the question. Some good products were developed (Pearson and Boruch 1986; Boruch and Pearson 1988). However, the SSRC conversations on how new statistical models and methods could be exploited to improve longitudinal surveys led nowhere.

One simple way to uncover answers to the question is to examine the writings of statisticians who invent new analysis methods. The presumptions are that these experts are in a good position to understand the implications of their work and, further, will have written about it. In the following section, we pursue this line of thinking and examine what appeared initially to be a promising approach and examine the published literature, proceedings, journals, and books.

Proceedings of the American Statistical Association

To understand what new analytic methods imply for survey design, it seems sensible to peruse the *Proceedings of the American Statistical Association: Survey Methods Section*. The 1993 edition was examined for papers describing new methods. These, in turn, were examined for a section on "Implications" or "Conclusions" that might educate us about answers to the question. We found none. (We did find implications in papers *other* than those dedicated to the mathematical invention.)

One might surmise that ordinary sessions of the ASA are usually not oriented toward the future. Rather, it may be more sensible to examine a source that is less time constrained, such as the *Proceedings of the Sesquicentennial Meeting of the American Statistical Association* (Gail and Johnson 1989). Boruch read each of the *Sesquicentennial* papers and looked for a sentence, paragraph, or section on implications and for conclusions that might have implications for the design of new surveys.

With a few exceptions, no paper in these special *Proceedings* directed attention to the matter. One of the exceptional papers described interviews with two able statisticians, Ron Gallant and John Pratt. The interviewer elicited their expert opinions about the implications of statistical theory for the design of a better census. Roughly speaking, both answered "I don't know."

Sending an e-mail inquiry to colleagues who are inventive about analytic methods seemed a sensible thing to do. So, a few of them were asked if they had *written* about the implications of their work for designing better surveys. Each individual had made remarkable contributions to analysis. Only one response is given here because it is instructive. It is from a colleague whom I admire on account of his inventiveness and industry.

> Thanks for your note. There's no doubt that better methods of analysis can lead to better designs. That idea permeates so much of what I do, I don't know exactly what to send. So I've decided to send you a CV and you can pick by interesting title. Also, I'll think harder to find particular appropriate articles.

We have depended on this scholar's work elsewhere in this report. His response reiterates the notion that implications of invention are important. But for able inventors, they cannot be drawn plainly, or will not be drawn plainly for many reasons, including the fact that "the idea permeates."

New Approaches to Analyzing Cohort Data: A Volume

Mason and Fienberg's edited volume (1985) handled advances in analyzing cohort data. The approaches to analyses are relevant to NCES surveys inasmuch as NCES sponsors surveys that attend to different cohorts of students in different time periods. Understanding the differences among cohorts and determining what may account for similarities or differences seems important. None of the papers in the Mason-Fienberg volume are explicit about how new analytic methods can be employed to improve any surveys, much less NCES efforts.

Failing to identify an explicit discussion of implications in Mason and Fienberg should not deter us, of course. Some implications may not be labeled as such. David Freedman's essay (1985) in the volume begins with the announcement that "[r]egression models have not been so useful in the social sciences" (p. 343). These models, for Freedman, include logics, time-series, and LISREL. His definition of social science includes education and psychology. His paper preceded recent developments in hierarchical linear models (HLM), but it seems reasonable to include HLM in his ambit.

Freedman argued that conventional statistical approaches to data analysis, as they are conventionally applied, have not had much yield. More important here, Freedman suggested that *any* new advances in statistical methods of analysis are likely to be uninteresting without major changes in the way that we think about data and about educational research and the behavioral and social sciences.

That is, the question posed earlier in this report, "Do new models and analytic methods have implications for better survey research design?" would have little merit for Freedman. It is the scientific thinking that underlies the models and methods that is important for him. Indeed, he argues that many of the models and methods are not sustained by good thinking about the processes that generate the observations in the first instance, i.e., a social theory.

It may not be difficult to agree with Freedman. Agreement, however, implies that the topic of this paper is misguided. Let us keep this implication in mind and resurrect it later.

The *Journal of Educational and Behavioral Statistics*: A Special Issue on Hierarchical Models

A recent issue of the *Journal of Educational and Behavioral Statistics* focused on hierarchical models (Kreft 1995). The issue's contents were reviewed to understand whether its authors suggested how surveys could be improved, based on advances in the subtechnology of hierarchical models. Only one author of an article in the journal stated that there are implications for design survey. His statements were opaque.

The Society of Industrial and Applied Mathematics

Curiosity and opportunity led us to ask about the topic of this report of a founding member of the Society for Industrial and Applied Mathematics (SIAM). SIAM's members, one might expect, would at times educe the implications of new analytic approaches in mathematics, including statistics for the better design of empirical research.

The interview with this scholar suggested that our mathematical colleagues are not inclined to speculate about how their work can be used to enhance future research. That is, mathematicians do not often educe the implications of their innovations for further work, at least not in print. The disinclination may, of course, be influenced by proprietary interests. Some members of SIAM are employed by profit-making corporations. University-based mathematicians who are also members of SIAM presumably have a taste for applied work. They invent new solutions to problems. But they also appear to infrequently educe the implications of their work and to make the implications plain in their published work.

The *Journal of Educational Statistics*: A Special Issue on Models

A special issue of the *Journal of Educational Statistics* (Shaffer 1992) reviewed the "Role of Models in Nonexperimental Social Science." David Freedman and Howard Wainer wrote their papers on structural models and on analyzing survey data, respectively. The commentaries and the authors' responses to criticism are important additions.

The authors did draw implications that bear at least indirectly on the design of some surveys, including perhaps NCES surveys. Freedman argued that "investigators need to think about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models" (p. 27).

In effect, this again suggests that we may have gotten off on the wrong foot in this report by focusing on the implications of new analytic methods. That is, for those of us who are interested in science, the theory ought to drive the way a model is built. The model, in turn, drives analysis: parameters that ought to be estimated, hypotheses that should be tested, and so on. This in turn can perhaps improve design of surveys, e.g., identifying assumptions whose tenability might be informed by certain designs and this leads to new models and analyses.

Wainer's conclusion was to "think hard" about nonresponse. In effect, this means inventing small theory whose elements might be informed by new data; he suggested that the new data are essential in understanding the nonresponse. Critics of the Freedman and Wainer papers argued along similar lines. Hope, for example, concluded "[t]here is no methodology that will write our theories for us" (p. 46).

To put the implications of these analysts bluntly: better theory (thinking) is warranted. This may not seem much like guidance for improving NCES surveys. But it does introduce some interesting choices for NCES that are discussed elsewhere in the report.

The Meaning of the Question at Hand

What was meant by "implications" at the outset of this essay was not made clear. Finding even a few implications reminds us to be more specific about what we seek. The word here means that, as a consequence of a new analysis approach, we might better understand any of the following (Exhibit 1):

1) What variables to measure or not to measure;
2) How to measure;
3) Whom to measure;
4) How many to sample;
5) When and with what frequency of measurement;
6) With what periodicity;
7) With what sample design characteristics (strata and so on);
8) In connection with what other data collection;
9) Why; and
10) How to report.

This list accords with at least some efforts to understand how to improve surveys generally. Items concerning what variable to measure, when, and on whom, are embodied, for example, in the products of a recent NAS-IOM workshop on integrating federal statistics on children (Board on Children and Families and Committee on National Statistics 1995). The list also accords with how users of new analytic approaches and data sets suggest improving survey design on the occasions that they do so, for example (Boe and Gilford 1992).

The list seems promising enough to use as a template for further work. Internet-based facilities that are discussed in the light of this report are suggested as a device for orderly acquisition of information on such items. Such a facility, a list server, for example, then provides a continuously updated archive of possible improvements based on the experience of users of NCES and other survey data.

The phrase "new analytic methods" as used in the title question may seem clear to some, but it is deceptive. Implicit in the phrase is the presumption that buried in any new method is a new model. A further presumption is that it is better to have explicit models to drive an analysis of data than to have analysis driven by implicit models. Both approaches are functional, however, to judge from the history of science including statistics. The former is regarded here as more functional.

Further, a new model might or might not have to depend on substantive scientific (educational) theory.

What can be regarded as "new," of course, is not obvious. Hierarchical models, though new to many users, are based on mathematical efforts that extend at least to Kempthorne and Cochran, and Cox in the 1940s and 1950s. So called network-based estimators are based in a fundamental way on elementary ideas about the probability of independent events. Spiraling methods now used in NAEP have their origins in balanced incomplete block designs developed over 30 years ago, and so on. The point is that here, when we denominate a method, model, or approach as new, the denomination is merely a convenient label.

And, of course, what a model or method *is* can be similarly complicated. Here, the focus is on a model that contains a stochastic error term and is suppose to represent reality—reality itself being partly represented by survey data. The models and methods examined here include hierarchical models, indirect estimators, design synthesis, and projection models, among others.

Published Analyses of Specific Data Sets

A search of education journals for 1991–95 uncovered 31 reports of analyses of data from NELS:88. Most of the authors employed conventional analytic methods such as OLS linear regression; perhaps three employed newer methods. Disregarding the analysis method, 15 out of 25 papers that we were able to review contained some form of implication. Nine articles contained no explicit statement of implications for better designing NELS:88.

Two papers were direct in providing very broad implications and indeed were developed to do so. These concerned the construction of math and science achievement tests so as to better recognize the multidimensional character of such ability. Of the 13 remaining papers, most called for new variables to be measured. Authors said that NCES should measure "global self-esteem" (instead of academic self-esteem), ask about criteria for placement of students into ability groups (instead of just asking whether students are grouped), ask how long students have lived in a single parent family (rather than just whether they do), elicit information on parental education and indicators of middle school philosophy (rather than just the existence of middle school). This list is idiosyncratic. That is, the implication drawn by the data analyst depends heavily on the analyst's particular theoretical framework and objective. This varies dramatically across analyses.

Only a couple of papers suggested that samples of certain groups be "beefed up," e.g., Hispanic students. And of course, some papers reiterated the need to collect similar data in the next wave of measurement, a tactic that NCES examines routinely.

This evidence suggests to us that some orderly way of identifying implications is warranted. The Terhanian Home Page model discussed later in this report is one option, a way of summarizing articles, implications, and analytic methods. It also implies that some method for routinely screening the published analyses is warranted; existing NCES advisory groups, for NELS:88 for instance, are an option.

Conferences, Working Groups, and Other Integrative Instruments

NCES and other federal agencies rely, from time to time, on specially convened groups to say something sensible about its activity. The group may be appointed by a department, as in the case of the NCES Advisory Council on Education Statistics, or the group may be appointed independently, as in the case of a National Research Council Committee. These and other groups might be expected to develop the implications of contemporary research for the future of the agency, including perhaps the design of specific studies. Some groups do so.

For example, researchers at the Educational Testing Service have occasionally tried to learn whether and how disparate databases that concern science could be used in combination. The Hilton (1992) effort, sponsored by the National Science Foundation, was unsuccessful in a few respects; it was successful in others. It employed rather than invented new methods or models. Surprisingly, Hilton's work (1992) dedicated little attention to how their lack of success could be rectified. That is, not much was said about how the design of independent surveys could be improved to foster their combination (see section on Linking NCES Surveys and Data From Other Sources).

Two other groups, which neither directly employ nor invent new statistical analyses, were also examined. Both dealt with the problem of "linking" data sets, the first being on teacher supply and demand (Boe and Gilford 1992) and the second concerning statistics on children (Board on Children and Families/Committee on National Statistics 1995). Both contain what amount to implications of prior empirical analyses and thinking, based on new methods and otherwise.

Teacher Supply, Demand, and Quality

Boe and Gilford's volume (1992) covers the NRC conference on this topic. Supported by NCES, the group was convened in the interest of enhancing the teaching force in the United States by focusing on major issues in the area and the information needed to understand them. This effort entailed reviews of the data that are produced, the data that might be produced, and the models that are used in forecasting supply or demand. The reviews perforce cover earlier analyses of the data, analyses that employ new methods or old.

This paper deals with "implications." In a sense, the NRC Conference on TSDQ also did so. It was "designed to reach a consensus . . . to stimulate suggestions concerning 1) information . . . and 2) further development of projection models and databases" (p. 3). The conference summary then provides NCES with another choice about how to characterize "implications." It and the main report are also interesting because they categorize the ideas/implications into two broad and arguably instructive categories: "information needs" and "suggestions." The needs usually refer to what variables ought to be measured. The suggestions focus on more specific implications. (Note that *none* of these are "recommendations"; the conference was not empowered to make them. This is a virtue in many respects.)

Exhibit 2 outlines the TSDQ Conference's summary of information needs. It is a short list of what variables ought to be measured by NCES and other communities of scholars even if we do not yet know how to measure them. We are told that we need, for instance, to measure teacher quality

(Information Need 1). The rationale is to better inform decisions about quality-quantity trade-offs and model-based forecasts of whether and how we might improve.

Implicit in some items is theory. Information need #3, for example, suggests that the demographic mix of teachers ought to be examined with respect to the demographic mix of students. Are old white people teaching young Hispanic people? What kinds of people are teaching whom? And, does it matter? Each question has an implicit, and rudimentary, theoretical basis. It seems important to recognize this basis and NCES can do so.

Some of the TSDQ Conference suggestions are outlined in Exhibit 3. Several points are worth noting. First, all the suggestions can be categorized using the generic list of implications in Exhibit 1, which reinforces the notion that this list may be a reasonable way to summarize such things. For example, Suggestions 1 and 2 bear on research to inform the use and measurement of the variable called teacher quality (Items 1 and 2 in Exhibit 1). Suggestions 18 and 19 bear on connections to other data sets, e.g., linking SASS to state databases bears on Item 8 in the list.

A second point worth noting is that the TSDQ Conference suggestions are a matter of collective judgement based partly on the expertise of participants and the papers commissioned for the conference. Backtracking to the volume's papers, we find most are based on rather simple but informative analyses. Murnane (1992), for instance, argued that state licensing records on teachers is a valuable resource and ought then to be linked somehow to the NCES effort, based on analyses showing downward trends in licensing and in licenses given to black college graduates and in their probability of returning to teaching having left the profession some time earlier, all from North Carolina records. Murnane also argued tersely for redesign of state record systems on account of the great difficulty he and his colleagues had in exploiting them. He did not recognize the NCES expertise in this area. But the crude implication we draw from this is that NCES' expertise on design of data systems and linkage is a major resource that might well be exploited in any effort to better capitalize on state data.

Only one paper in the TSDQ Conference *Proceedings* focused on models, and using them in the context of NCES surveys and state data. Barro's concerns (1992) lay solely with projection models of different kinds and the data used to sustain their use. His paper is nonetheless instructive because of the implications that were drawn from it by Boe and Gilford (1992), such as Suggestions 18 and 19, and on account of Barro's own thinking. Indeed, Barro's entire paper can be regarded as an exercise in drawing implications. For instance, he argued that the mechanical (demographic) demand models in contemporary use are far less informative than new behavioral models that help one address "what if" questions. His implication is that NCES ought to use the "what if" theme to drive design; NCES' current projection models are of this variety (Gerald and Hussar 1992). Improvements, according to Barro, lie partly in adding variables such as pupil-population ratio and teacher salaries. It lies partly in treating a measured variable, notably state aid to schools, *not* as exogenous but as a variable that itself ought to be forecast from other (unspecified variables). Other suggestions lie in frequency of measurement (Item 5 in the generic list); more being better, in using state-level data to build more detailed and policy-relevant models (Item 8).

The idea he produced for better designs based on the supply side are sustained by simple rather than elaborate models and findings from their application. His implications are numerous. Among other things, he reiterates the need to exploit SASS to get better forecasts of teacher attrition rate, especially 1-year followups of subsamples to get at turnover.

Many of the implications that Barro drew seem important. They are certainly ample. About one implication appeared every page and a half in the discussion on demand. One implication that we draw from the way Barro approached his task and the TSDQ Conference *Proceedings* is, again, that the generic list in Exhibit 1 is helpful in classifying implications. A second, more important, concern is the mechanisms available to NCES to uncover implications on new *or* old models and their application. A conference was organized to do so. Third, when implications are ample, we need to keep track of them and their bases. The generic list in Exhibit 1 helps the orderly acquisition. Sharing such information beyond print would arguably help (see the section on New Technology).

Integrating

The Board on Children and the Families and the Committee on National Statistics (1995) of the NRC/IOM convened a workshop "to examine the adequacy of federal statistics on children and families" (p. 1). Its joint sponsorship, by the Board and the Committee, and the topic itself led us to expect "implications" to be produced and indeed they were. The final report, *Integrating Federal Statistics on Children* (hereafter called *Integrating*), is plentiful in its supply of them.

The summary of *Integrating* outlines cross-cutting "suggestions" (p. 2) based on collective expertise and commissioned papers, as in the Boe and Gilford (1992) effort. But the summary is rather broader in its handling of them. We are told the following, for example:

> Improvements in data are needed to understand the connections between resources and child outcomes, as well as family and community processes that translate resources into outcomes (p. 3).

This is rationalized by recognizing the availability of data on input variables (e.g., PSID) and offering the opinion that "data on child outcomes are substantially more limited" (p. 3). This "implication" does not recognize, much less exploit, the notion that children's education achievement is an outcome, that NCES routinely obtains such information *and* information that bears on some resources. Two sub-implications were drawn: that data ought to be collected for "more than purely descriptive purposes . . . " and that the use of time by parents and children is a major variable that is rarely measured. The first item is relevant to NCES in that the agency is often confronted by the need to incorporate substantive theory into debates about design of surveys. The second is relevant inasmuch as NCES has asked questions about how time is spent in some surveys, e.g., time on teaching certain topics and time in watching TV. Again, this is unrecognized in *Integrating*.

Integrating's summary is about what variables to measure, as in the example above; about family relationships (e.g., biological, adoptive, step, and noncustodial parents); about the need for service-related data at subnational levels; about new strategies (designs) for oversampling certain groups; about "improved longitudinal data . . . to address . . . policy issues [on] changes in family resources, predictors of successful development . . . precursors of serious problems; . . . " and about cross-agency planning and coordination.

Exhibit 1 catalogs the implications that are drawn in *Integrating's* summary. The crude enumeration suggests that the generic list of implications developed earlier seems reasonable. It is important to note that the implications are not drawn directly from new analytic methods nor are they drawn specifically from *any* method. They are drawn in unspecified ways from various and often unspecified analyses. In other words, the coupling between "implication" and analyses is often loosely specified. Brooks-Gunn, Brown, Duncan, and Moore (1995), for example, recognized that hierarchical models can be employed to analyze NELS:88 on account of this survey's design (p. 63).

Let us backtrack to the papers that were written for *Integrating* to understand more specific implications for NCES surveys. What do we learn? First, Brooks-Gunn et al. (1995) admired NELS:88 for the survey's attention to eliciting information from multiple sources, such as parents, teachers, children, and school administrators to produce data that help us to understand outcomes and inputs and process overtime. The only implications drawn by Brooks-Gunn et al. are that 1) the 1996 wave of measurement of children who would be in the 20-24 age range ought to be done; and 2) NELS:88 ought to be continued until the cohort is at the age of 28 or so (in the year 2003). The rationale for the authors lies in their view that transitions, from late adolescence to adulthood for example, are important. It is not based on identified data analyses or particular analytic models or methods (p. 76).

Hoffreth's paper (1995) in the same volume focuses on transitions to school. She then emphasizes the need for an entirely new survey, the Early Childhood Longitudinal Survey, that has been considered by NCES. The rationale is that we know less about entry to schools than we should. Further (p. 114), the United States has no longitudinal study underway that begins prior to entry to school. Hoffreth was attentive to linkage among data collection efforts but did not mention NCES in this context nor did she get much beyond the notion that data on mothers from the NLSY ought to be coupled with other data.

The Implications of Looking for Implications

This primitive review of scholarly published works that might have contained implications itself has implications, of course. The zero point implication is that the question posed at the outset of this essay was not put quite rightly. That is, getting beyond the initial question is important. We cannot be content with: "What do new analytic methods imply for NCES survey design?" We must ask the further question, "What are the implications of employing new analyses or old ones for design of NCES surveys?" Also, what do we mean by implications? And who articulates them?

First, we should not expect able scholars who *invent* new methods of analysis to educe and state plainly the implications of their work for designing better surveys. Attention to such implications is sparse in the current culture of mathematical statistics.

Second, we should expect fewer than half of the scholars who apply new or old methods to real NCES data to make suggestions (implications) about improving survey design. Further, we should expect them to suggest: new variables or deeper/more sophisticated measurement of existing variables. The need to oversample certain groups or to measure the same way is, at times, reiterated.

Third, when the implications are stated at all, they are diverse and depend heavily on the analyst's idiosyncratic interests and theoretical perspective. When the stated implications are unclear, as some are, they can be perfectly uninformative and may require further action. The diversity means that NCES might develop methods for orderly acquisition and screening using vehicles that NCES has at its disposal (the Web, advisory groups, and so on). That is, many implications, can be generated and this is another peculiar problem. Some options for handling the problem via the Internet are described in the last section of this report.

Fourth, mechanisms exist to foster statements about implications of new analytic methods of employing new or old analytic methods to NCES data. The mechanisms include institutions such as NAS. They include grants, e.g., the Stanford group. They include professional organizations and journals to which NCES professionals contribute pro bono. NCES can encourage its contractors to educe the implications of their work for improving survey designs and can influence grant agencies to encourage grantees to educe the survey design implications in their research.

EXHIBIT 1

THE POSSIBLE DESIGN IMPLICATIONS OF A PARTICULAR METHOD, MODEL, OR ANALYSIS*

1) *What* new variables should be measured and what variable ought not be measured?
2) *How* or what should we measure?
3) *Whom* to measure?
4) How many?
5) When and with what *frequency*?
6) With what *periodicity*?
7) With what broad *design* (e.g., strata, and so on)?
8) In *connection/coordination/link* with what other survey, database, or experiment?
9) How to *report*?
10) *Why* for each of the above?

---

*For example, Mullis, Jenkins, and Johnson's HM analyses of NAEP data (1994) suggest that NAEP should better measure sets (Item 2), more instructional variables ought to be measured (Item 1), and information ought to be elicited from more teachers (Items 3, 4) in the interest of understanding the relative effects of classroom/teacher (Item 10).

EXHIBIT 2

INFORMATION NEEDS: TEACHER SUPPLY, DEMAND, AND QUALITY*

Information Need 1:	Teacher quality indicators
Information Need 2:	Teacher credentials
Information Need 3:	Demographic matching
Information Need 4:	Teacher professionalism
Information Need 5:	Programs to improve practice
Information Need 6:	Assessment of quality of teaching practice

---

*Excerpted from Boe and Gilford (1992).

EXHIBIT 3

SUGGESTIONS: TEACHER SUPPLY, DEMAND, AND QUALITY*

Suggestion 1: Teacher quality indications; Sustained research

Suggestion 2: Tested ability of teachers; Tests of knowledge

Suggestion 8: Reserve pool; Little is known; Survey applicants in SASS

Suggestion 16: Teacher demand data; NCES should develop a (better) model for teacher demand projections

Suggestion 18: Unused databases (e.g., NSY and Supply)

Suggestion 19: Linking SASS and state DBS

Suggestion 23: TSDQ Consortium

---

*Excerpted from Boe and Gilford (1992).

EXHIBIT 4

THE IMPLICATIONS DRAWN BY THE NRC GROUP ON TEACHER SUPPLY, AND SO ON (TSDQ)
THE NRC GROUP ON INTEGRATING FEDERAL STATISTICS*

		TSDQ	Summary Integrating
1)	Variables: New/deleted	Yes/no	Yes/no
2)	Measurement	Yes	
3)	Sample units		Yes
4)	Sample size: Increase/decrease	Yes/no	
5)	Time		
6)	Timing		Yes
7)	Survey design	Yes	Yes
8)	Links	Yes	Yes
9)	Reports		
10)	Rationale	Yes	Yes/no

---

*For example, *Integrating Federal Statistics on Children* (Board 1995) educes implications from other research for the design of new surveys. The implications cover sample size (e.g., oversampling Hispanics) and links (e.g., to state databases). Some implications bear on NCES efforts and they are identified by a "yes" in the column "Summary Integrating."

CROSS-DESIGN SYNTHESIS: IMPLICATIONS FOR THE DESIGN OF EDUCATIONAL SURVEYS AND CONTROLLED FIELD EXPERIMENTS

Background: Cross-Design Synthesis

Cross-design synthesis is a strategy for combining analyses of the data that are generated in controlled experiments with analyses of data generated from surveys or from certain administrative databases. For example, the national data obtained in a NCES probability sample survey on adult literacy in the United States might be used in an analysis that purports to yield estimates of the relative effects of certain literacy programs. The results would then be combined with evidence generated by a dozen experiments on the relative effectiveness of local literacy programs.

The object of this combination of evidence is to produce valid and generalizable estimates of the effect on certain social programs. The rationale for combining the different data sources is that the combination exploits a benefit of controlled tests, notably an unbiased estimate of the treatment effect in local settings, and further exploits a benefit of national probability sample surveys of the kind that NCES executes, the capacity to make generalizations to a larger target population.

In the adult literacy case, controlled experiments in particular sites may yield valid estimates of the effect of literacy programs. But the estimates are local, e.g., of uncertain generalizability. The national database or survey may yield estimates of the effect of programs at the national level. These latter estimates are suspect in that their validity is unclear; the survey or administrative database involves no active control. Rather, analysis usually involves statistical control. A combination of the two sources of evidence might be combined so as to justify inferences that are both valid and generalizable.

The general approach to cross-design synthesis is described in a U.S. General Accounting Office report (USGAO 1992) and in Droitcour, Silberman, and Chelimsky (1993). A more recent report (USGAO 1995) describes the approach's application to the problem of estimating the effect of breast conservation versus mastectomy on the 5-year survival rates of women with breast cancer. This analysis is based on data from randomized clinical trials and a large database. In particular, six studies serve as the evidence in the randomized trial category; they include single-site and multisite experiments undertaken in North America and Europe. The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) system constitutes the administrative database. It provides data on breast cancer patients, their treatment, and prognosis based on reports from practicing physicians in a large geographic region of the United States.

Objective and Assumptions

Recent reports on cross-design synthesis have focused on the analysis of data from two kinds of sources: controlled experiments and databases. Here, the focus is on how the thinking about cross-design synthesis can improve the design of administrative databases and national surveys sponsored by NCES.

To put the objective bluntly, we want to turn "cross-design synthesis" into a vehicle for better design of studies, rather than to encourage its current use as a form of meta-analysis. This objective accords with the theme of this report, i.e., educing the implications of new analytic approaches for better study design. It is also distinctive; the inventors of this analytic approach did not develop this implication (Droitcour and Chelimsky 1995; Boruch 1995).

A first assumption is that it is important to estimate the effects of education programs in the United States. The second assumption is that NCES cannot or should not undertake formal evaluations of the effects of such programs. Other federal agencies, for example, are responsible for running controlled experiments on education programs. Third, we assume that NCES can design surveys that accord with the first two assumptions. Finally, we assume that, in 5 years, we will have to combine results from different sources to reach a conclusion about a program's effectiveness.

The object here is to address the question: How can NCES exploit ideas in the cross-design synthesis approach so as to design better surveys or databases?

Definitions

Survey here means an effort to elicit information from a probability sample of individuals or institutions who are members of (ideally) a well-defined target population. Such a survey involves no active treatment or manipulation of respondents, apart from the act of eliciting information. The survey may be cross-sectional, for example, the NCES 1991 National Adult Literacy Survey. Or, the survey may be longitudinal, as in the case of the National Educational Longitudinal Study undertaken in 1988 (NELS:88).

Administrative database here is defined as a set of administrative records on a well- defined target population. For instance, transcripts on all students in a junior college, containing information about the students' courses and grades, constitute a database. The records on all students in a voluntary service organization's program on literacy also constitute a "database."

A database is a survey of a special kind. It usually includes the entire target population; no sample is taken. It is a "survey" to the extent that any set of administrative records is the product of interviews of a kind that are often done in survey research, albeit under different conditions.

Certain kinds of NCES data collections result in a database of administrative records for eligible institutions in a population. That is, the NCES effort is not based on a sample. The population databases include the Common Core of Data (CCD), the Integrated Postsecondary Education Data System (IPEDS), and the Library Statistics Program. In principle, analyses based on data from any of these sources could be combined with results of controlled experiments in a cross-design synthesis.

A controlled field experiment is a setting in which individuals (or other entities) are assigned to program variations in accord with a plan designed to produce an unbiased estimate of the differences among the program variations and a statistical statement bearing on one's certainty about the results. For instance, one may design a study to compare certain approaches to teaching English as a second language so as to understand which approach works best, and under what conditions. Individuals or entire organizations might then be randomly assigned to the different program approaches, engaged in the relevant approach, and then measured with respect to their English proficiency.

Because controlled experiments are difficult to mount, only a few are undertaken in a very small number of sites. The results may be relatively unequivocal in the sense that one variation appears to work better than another in one or more of the sites. It will usually not be clear how these results can be generalized. For instance, the experiment sites may include cities in the Northeast; they may exclude the Northwest and Southwest.

An agency such as NCES is mandated to conduct observational surveys. It is not mandated to execute controlled tests of education programs. Other agencies within the U.S. Department of Education, such as the Planning and Evaluation Service, are mandated to conduct controlled experiments to evaluate education programs. Further, private foundations and other government agencies may exploit surveys or experiments or databases to further knowledge about programs or about the educational state of the nation.

Rationale in the NCES Context

The first rationale for focusing on cross-design synthesis is as follows: Users of NCES survey data have often tried to use the data to estimate the relative effectiveness of different sorts of education programs. It seems reasonable to expect these efforts to continue despite the ambiguity in the interpretation of the data that is bound to occur because the survey is a passive instrument rather than an active experiment. Insofar as cross-design synthesis carries a promise to combine such survey data with other data from experiments, so as to produce better information, it is sensible for NCES to exploit opportunities presented by cross-design synthesis.

A second rationale is more ambitious. It is that cross-design synthesis can be a vehicle for the mutual education of survey researchers and experimenters and a productive change in scientific culture. Thoughtful survey researchers cannot always be well informed about controlled field experiments. For example, Clifford Clogg (1989), a sociologist and survey statistician, announced that "experimentation of the classical variety is usually impossible, inconceivable, or difficult to implement." Economists and educational researchers, such as Henry Levin, and mathematical economists, such as James Heckman, who rely heavily on observational survey of that sort that NCES produces, have made similar claims. They rarely present empirical evidence (see Boruch 1994 and references therein).

Experimenters, on the other hand, depend in only a limited way on survey data of the kind that NCES obtains. Their design of a local controlled experiment on the relative effectiveness of two compensatory literacy programs may, for example, depend on regional or state literacy rates to inform the experiment's design. As a consequence, experimenters are at times not well informed about surveys run by NCES or other statistical agencies. Few important controlled experiments in the United States rely heavily on surveys run by federal statistical agencies except at the experiment's design stage, where the experiment and design may recognize survey-based estimates of the incidence of a problem.

A More General Rationale: Government Agencies

A broader reason for inverting the analytic idea of cross-design synthesis so as to focus on design of surveys is that the approach can be a fine bridge between the members of the federal statistics agencies on the one hand and the federal evaluative agencies and private foundations that sponsor controlled experiments on the other. These include, for instance, the Bureau of Justice Statistics responsible for the National Crime Victimization Surveys, and its sister agency, the National Institute of Justice which is responsible for multisite controlled experiments on the police handling of domestic violence, among other topics. It includes the Bureau of Labor Statistics, an agency that continues to run large-scale probability sample surveys on employment and training and the Department of Labor's unit for large-scale experiments on residential Job Corps, the Job Training Partnership, and others. The role of NCES as statistical agency is complemented by the role of the Planning and Evaluation Service at the Office of the Undersecretary at USDE.

The gap between the statistical agencies and the other units that focus on analysis represents a kind of intellectual travesty in this country, given that data from the former *are* often used to estimate program effects, not just to describe them. The insulation of statistical agencies such as NCES has considerable political justification, of course. Statistical data should be and, under current laws, is relatively free of political influences. Analysis units are more vulnerable to the latter although some have a fine reputation for both independence and political sensitivity. The institutions need to keep the two functions separate. But this does not vitiate the idea that as an intellectual matter, the separation is unnecessary and arguably dysfunctional.

The gap between the statistical agencies and those responsible for analytic studies of programs was recognized implicitly and explicitly in a NRC volume on integrating statistics on children. Brooks-Gunn et al. (1995) and Hoffreth (1995), for instance, recognized the distinctive role of the JOBS experiments and the Perry Pre-School Project in the context of NCES and other surveys but did not explore the matter deeply. Pallas (1995, p. 153) recognized the merits of NCES and other statistical systems and the distinctive role of experiments on dropout prevention programs, and more importantly, expressed discomfort with the volume's heavy emphasis on statistical systems. It is a discomfort that we share, discussed briefly in a paper on the future of experiments (Boruch 1994), and explore here.

The First Illustration in the NCES Context

The NCES has undertaken a national probability sample survey of adult literacy in the United States with augmentation for special subpopulations, e.g., prisoners. Reports on adult literacy are available from Andrew Kolstad's Education Assessment Division at NCES (see Davis and Sonnenberg [1995] and other NCES *Programs and Plans*). Suppose that the NCES will run another such survey and that the survey's plan can be influenced.

The U.S. Department of Education's Planning and Evaluation Service, Office of the Undersecretary, has had a responsibility for evaluating the effectiveness of certain adult literacy programs. Suppose that another evaluation at multiple sites will be undertaken by this office.

Regard the NCES survey on adult literacy and any other information obtained by NCES from administrative sources as a database. Regard the USDE/PES evaluation as a source of data generated by controlled experiments.

Consider then the question: How can the cross-design synthesis approach inform the design of new surveys or databases (and experiments) in the adult literacy arena so as to generate better estimates of the effect of literacy programs in 5 years?

The GAO reports on cross-design synthesis approach suggest that in the survey and in controlled experiments we attend to the following:

- Target population and its characteristics;
- Treatments;
- Outcomes; and
- Propensity scores.

Each is considered in the section that follows.

Implication: Target Population and Samples

Cross-design synthesis requires that the individuals who are targeted in controlled field experiments are also represented in the survey sample or database.

A new NCES sample survey on adult literacy in the United States must then include individuals who are targeted for literacy services. Attempts to estimate the effect of the services, undertaken in local controlled experiments, must target similar individuals.

For instance, if programs make major efforts to serve illiterate immigrants from Bosnia, Slovakia, Morocco, or other countries, then NCES must plan to include these in the target population for a new NCES survey. This, in turn, requires that the local literacy agencies be able to specify their main local targets. It implies that the federal agency responsible for support of adult literacy programs, an agency different from NCES, be able to specify target population that is of major interest in any controlled experiments that are undertaken to test the programs.

Implication: Treatments

To combine data in the cross-design synthesis approach, one must know what treatments (programs) are delivered to whom and when. A new sample survey of literacy in the adult population undertaken by NCES then would have to ask individuals about the literacy programs in which they have participated. Learning *how* to ask such a question so as to secure reliable responses is difficult, to be sure. Figuring out how to exploit local databases of literacy services that maintain such information is also likely to be difficult. Nonetheless, NCES must do so if the object is to produce a cross-design synthesis in 5 years, of who gets what literacy program and to what effect.

For a federal agency or private foundation that sponsors controlled experiments on the effects of certain literacy programs, the implication is that the agency or foundation must record the individual's program participation. More important, the method of recording must correspond with how the NCES national survey asks about program participation. Questions about program participation are framed in a survey and the way they are framed in local experiments must be compatible with one another. The local experiments will usually depend on administrative program records to establish an individual's participation in a certain program. A survey usually involves depending on an individual's self-report about participation in a program; it may also depend on institutional records contained in databases.

To make the two kinds of information compatible for cross-design synthesis, several options might be considered. The local experiments might ask about participation in the same way that the survey asks, permitting one to correlate self-reports with administrative records. Or, both the survey and the experiments might direct attention to local service providers and their clients, eliciting records so as to reduce reliance on self-reports of individuals. In any case, small studies of the matter are needed.

Implication: Outcomes

The impact of adult literacy programs can be registered partly by measuring an outcome variable such as "literacy level" of each individual or of groups of individuals.

To accomplish a cross-design synthesis of the effects of literacy programs, a survey agency such as NCES must cooperate with an evaluation agency such as USDE/PES or a private foundation that sponsors evaluations in developing outcome measures. That is, the organizations must agree on how literacy level is to be measured.

Cooperation of this sort is not easy across local literacy programs, much less across federal agencies or private foundations. For instance, a recurring problem is that local literacy programs, regardless of their sponsorship, have not been able to agree on how to measure literacy. In the absence of agreement, no surveys or experiments undertaken by the federal government are likely to lead to a persuasive cross-design synthesis of whether and which programs work in what sense.

Implication: Propensity Scores

A controlled randomized experiment relies on randomization to produce an unbiased estimate of the difference between two or more groups. In such an experiment, individuals who are eligible to be served by a literacy program and who are willing to avail themselves of the program are randomly assigned to the program or to one of two or more variations of the program. Or, entire organizations might be allocated randomly to alternative service programs. In ordinary language, the groups are "equivalent" apart from chance because they were randomly composed. A comparison of the groups' performance is then fair. The difference in average literacy level of the two groups following their engagement in the programs, or difference in rates of achievement then provides a good estimate of the relative effectiveness of the program variations.

The NCES does not sponsor controlled randomized tests of literacy policies or programs. NCES does, however, provide an observational survey data platform for estimating effects. Statistical analysts who rely on such a platform have usually developed strategies to approximate the results of a controlled experiment, i.e., compensate for the absence of the randomized test. The strategies vary. During the 1960s, for example, analysts employed OLS estimates of a program effect that was based on a simple, single-stage linear model and observational data (e.g., covariance adjustment).

The focus here is on propensity scores as a device to produce analyses that approximate the results of a controlled test. Such scores were used, apparently to good effect, in the GAO (1995 and Appendix I) report on the differences between two approaches to treatment of breast cancer. The recent work on propensity scores has the benefit of conscientious thinking about how to recognize the fact that people, in ordinary circumstances, do not engage in programs randomly, and how to incorporate this and related selection factors into analysis.

The GAO's application of cross-design synthesis to data on treatment of breast cancer suggested the following were important in developing propensity scores:

1) Year at which the individual is engaged in treatment;
2) Geographic area of residence;
3) Severity of the problem at baseline;
4) Age of the individual;
5) Race or ethnicity; and
6) Marital status.

How and why the variables were chosen is not made plain in the GAO's report (1995).

These same variables *seem* relevant nonetheless to understanding the propensity of individuals to engage in adult literacy programs. The access to such programs was greater in 1990 then it was in 1980, and the efforts to entrain clients has arguably been more vigorous in the past few years. Year of engagement then is arguably important. The geographic area of residence and ethnicity are related and theorists argue that it is important to recognize each. For example, Hmong immigrants have clustered in only a few cities in the west, midwest, and northeast United States. Bosnian immigrants and others from the new independent states of the former USSR make their homes elsewhere.

Marital status may have no obvious influence on one's inclination to become literate. But a conscientious theorist might argue that if one examines the way families develop once marriage occurs, the way adults in the family behave in their children's interest and in their own economic interest, the variable called "marital status" may be a reasonable one to use in constructing a propensity score.

Implication: Propensity Scores, Intentions, and Reasons

Roughly speaking, a propensity score reflects the predilection of individuals to belong to one group rather than another, where the predilection is indicated by some observable characteristics of the individual. More specifically, it is the conditional probability of being in a particular group given a vector of observed covariates (Rosenbaum and Rubin 1983).

For example, high school dropouts and high school stayers constitute two groups. The probability of being in one group or the other can be characterized descriptively as a function of variables such as daily school attendance rates, age, academic grades, and plans for higher education. Similarly, the probability of entry to college or the work force can be characterized as a function of demographic and other variables.

The variables typically used to estimate a propensity score usually include demographic and contextual information. Over 30 such variables were used by Rosenbaum (1986) to estimate a kind of propensity score for school dropouts and stayers. They included those identified in the paragraph above.

The variables used to compute a propensity score are often "indirect" in the sense that they indicate an individual's state, rather than capturing directly: 1) an intention to belong to one group or another, or 2) the observable reasons for belonging to one group or another. Education surveys, with a few important exceptions, do not ask individuals why they dropped out of school or about their intentions to do so.

An implication of the analytic work on propensity scores (and related analytic methods) is that we should consider obtaining information on the individual's intention or on the reasons for membership in a group or both. One rationale for obtaining such information is that it *appears* to be a more direct covariate of membership than less direct ones, such as demographic characteristics. The connection between an individual's declaring that he or she will drop out of school and actually doing so appears more direct, less distant, from actual membership in the dropout group (i.e., becoming a dropout) than, say, the connectedness between "age" in school at one point in time and becoming a dropout in another.

Usually, no formal educational theory underlies the construction of propensity scores. Rather, the justification for their use lies in small and large sample statistical theory (Rosenbaum and Rubin 1983). A second rationale for eliciting information about intentions or reasons then lies in the need to construct better substantive theory in education. To the extent that the propensity approach can be informed by education theory and can help build the theory in a cyclic way, this seems desirable. Better theory, for example, may promote propensity scores that are easier to compute or more interpretable. They may decrease the need for a large reservoir of cases on which to match when propensity scores are used with matching. This promotion may hinge on eliciting information about intentions or reasons.

Sensible readers can quarrel with the idea that information about reasons or intentions ought to be elicited in surveys. Critics do so with considerable justification. Asking individuals about intentions and reasons is difficult and, in any case, may not be useful. For instance, Rosenbaum's exploration (1986) of a propensity-scorelike approach in a dropout study using NCES' High School and Beyond data uncovered the fact that "the vast majority of students who eventually dropped out said in their sophomore year that they expected to graduate" (p. 208). Was the question asked well? We do not know. We do know that other "intentions" question, about aspirations beyond high school, was indeed useful to Rosenbaum in constructing the propensity score.

At least some scholars would argue, based on good evidence, that the more general problem is of understanding revealed preferences and their usefulness in studies based on observational data. Manski's book (1995) has a chapter dedicated to this and related matters. The NCES' National Longitudinal Study of the High School Class of 1972 (NLS-72) served as a vehicle for his attempts to understand how college enrollment rates would be affected by Pell Grants to needful students. The variables he used as a surrogate for revealed preference included ability, income, and so on as measured in NLS-72.

Recognizing the skepticism that economists have about self-reported preferences, Manski argued persuasively for trying to measure the preferences directly. Part of the argument is tied to theory, notably theory about what variables to use in an analysis. Economists vary, for example, in the variables they have included in studies of returns to schooling (p. 97). Manski's argument is based partly on empirical grounds. He provides citations to research in the arenas of consumer buying intentions, fertility (based on Current Population Survey over the last 50 years), and voting intentions, and to work by social psychologists in the arena to justify his argument that preferences ought to be assessed more directly.

For Manski, one of the implications of agreeing that information on preferences is important in that we must get beyond simple "yes" and "no" answers, e.g., "Do you think you will drop out of school?" He argues, on analytic and empirical grounds, for eliciting a probabilistic assessment of behavior from each individual. To paraphrase his sample question: "Looking ahead, what percent is the chance that you will drop out?" Social psychologists working in the arena would probably go further to argue for eliciting preferences (self-predictions) at points in time that are close to the event in question. Asking in September about students' perceived probability of dropping out is arguably less useful than asking the question in November or December.

To summarize, propensity score approaches suggest that 1) we consider more seriously whether to measure preference (self-declared propensity), and 2) how and when the preferences are measured seems important. But we need to do research on this.

Similarly, one may argue that to do a better job constructing propensity scores, one ought to observe or elicit information on why or how people find their way into groups, e.g., into a literacy program or not. To return to the main illustrative context, NCES might then ask a question of the following sort: "Which of the following factors influenced your decision to enroll (or not enroll) in the literacy program?"

The responses to the question might then be incorporated into a propensity score that is better than (say) one that relies solely on demographic information. Further, the responses may help to develop a small part of a substantive education theory that helps to understand processes by which people enter programs or, more generally, a substantive theory that complements or augments statistical theory for analysis of observational data.

A question of the sort proposed above appears not to have been asked in any large-scale observational surveys, nor can we find concrete illustrations in the published reports on selection modeling or propensity scores (e.g., Rosenbaum and Rubin 1983; Rosenbaum and Rubin 1984; Rosenbaum 1989). Ways to frame such a question can be developed, based perhaps on NCES expertise and cognitive research in a laboratory or field setting.

Implication: Measurement Issues

In national probability sample surveys, we can often measure a variable using only one or two questions or using an inventory with very few items. Learning about children's relations with other children in a survey might, for example, involve only a few questions about (say) how many friends that the child says he or she has. A set of local experiments designed to test ways to improve the ability of withdrawn or hostile children to relate to other children usually involves a more elaborate inventory. It is not clear how to link the data from sparse measures made in a large sample survey to the deeper measures made in the small sample experiments.

Similarly, learning about literacy level of individuals in a large sample survey must contend with respondent burden. Local experiments can often depend on inventories that demand more time of the individuals who participate, and do.

The problem here has a delicious analogue in atmospheric weather research. Satellite imaging might be based on measures on grids that are 1,000 kilometers in width. Surface measures may be obtained in far smaller grids, 100 kilometers across for example, yielding more precise local measurement. The challenge lies partly in how to integrate these data across levels of resolution (Draper et al. 1992).

Learning how to measure simply in large-sample surveys and how to measure roughly the same construct with more precision in local experiments are important. Cross-design synthesis and the problem of combining different sources of information generally, invites us to learn how to link the two sources.

Summary

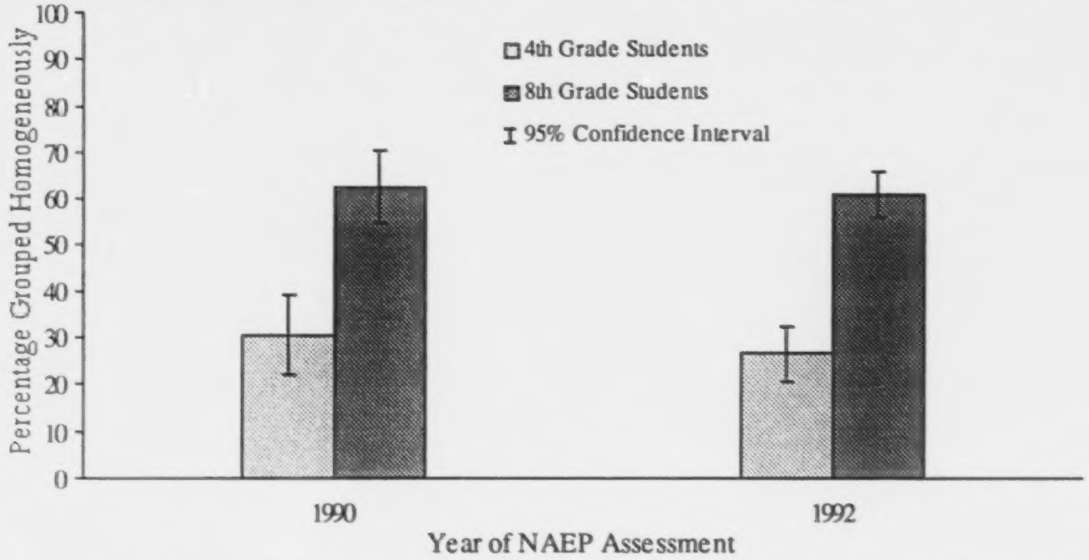
NCES has taken a leadership role in arenas related to cross-design synthesis. This strength suggests that it can succeed in work based on design orientation to cross-design synthesis. For example, the Common Core of Data (CCD) is a substantial product of NCES' efforts at the national level to develop agreement among the states about what ought to be measured and how what is measured ought to be defined so that resultant data are interpretable.

This achievement is no small one. The experience in negotiating with different jurisdictions and the product are valuable. Both can be capitalized in exploiting a cross-design synthesis approach.

A Second Illustration: Ability Grouping

Schools often sort students by ability (homogeneously) in math classes, particularly in higher grades. As Figure 1 shows, nationally representative data indicate that schools[1] grouped a significantly[2] higher percentage of 8th grade students than 4th grade students by ability in 1990 and 1992.

Figure 1—Percentage of homogeneously-grouped 4th and 8th grade public school math students in 1990 and 1992



SOURCE: U.S. Department of Education, National Center for Education Statistics, *NAEP Data on Disk: 1992 Almanac Viewer*.

Numerous propositions that attempt to explain why ability grouping increases in higher grades seem plausible. Students of mixed ability, for example, may receive academic instruction in several subjects from one teacher in elementary school (i.e., K-5), reducing the possibility of homogeneous grouping. As students reach middle school (i.e., 6-8), however, they may receive instruction in several subjects from several teachers. It may then become more convenient to group by ability; that is, to reorganize heterogeneous groups of students into homogeneous ones. Or differences in achievement may accumulate as students age, becoming more pronounced in later grades, thereby creating the perceived need for homogeneous grouping. Or schools may intentionally or otherwise sort students by socioeconomic status, gender, and race, as some critics of ability grouping have charged. Or decision makers may believe (perhaps on the basis of research evidence) that comparable students, particularly older ones, learn better in homogeneous classes. Hereafter, this paper will attend primarily to the latter two propositions.

The Concerns of Equal Opportunity Advocates

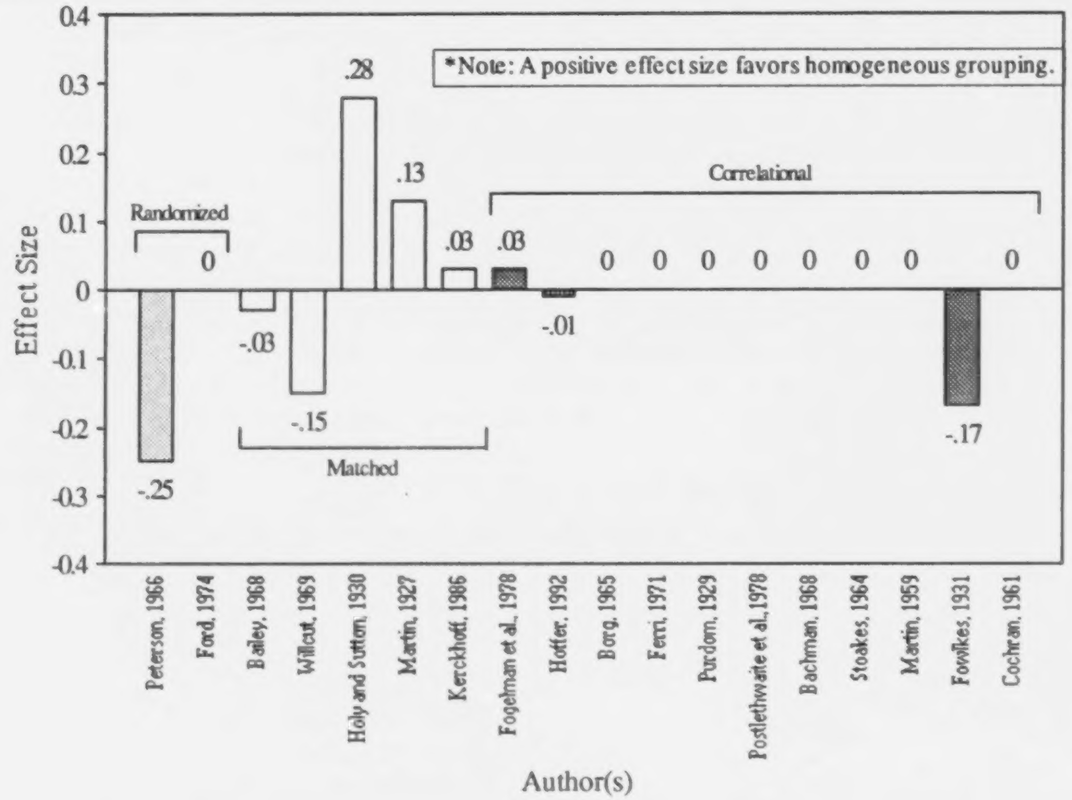
Advocates for equal opportunity often assert that two tracks—one leading to prosperity and the other to poverty—exist in America's schools. That these tracks appear to reflect gender, racial, and socioeconomic differences is cause for alarm. "As a result of the two track system," Beatrix Hamburg, president of the William T. Grant Foundation, writes: "[T]here is educational neglect and underachievement that disproportionately afflicts girls, minorities, and the poor" (1993, p. 9). And "what purpose has desegregation served," Jay Heubert, an attorney and education professor at Harvard University, adds "if resegregation takes place within desegregated schools?" (personal communication, November 1992). Ability grouping, from their collective perspective, may be viewed as one vehicle through which differences along gender, racial, and socioeconomic lines are bred and perpetuated. And indeed, some evidence supports this view. Oakes (1990), for instance, has observed that

- Schools tend to disproportionately place black, non-Hispanic and Hispanic students in lower ability groups;
- Ability groups tend to reflect socioeconomic status;
- Teachers of low ability groups tend to expose students to fewer, less demanding, topics than do teachers of high ability groups; and
- Schools tend to place their least qualified teachers in low ability classes and their most qualified teachers in high ability classes.

Researchers Concerned with Student Achievement

Those concerned with student achievement, meanwhile, often assert that ability grouping either impedes or adds no value to overall math achievement. Understanding whether this is so suggests the use of experimentation. In question form: If a sample of students were randomly assigned to homogeneous and heterogeneous instructional groups, which group would achieve at a higher level? Asking the question is the easy part. Mounting randomized experiments has turned out to be more difficult—there have been none since 1974—and there are only two on record. But there have been several non-randomized (i.e., "matched" and "correlational," in Slavin's terms [1993]) efforts to estimate the impact of homogeneous grouping on math achievement. Slavin (1993) included 16 such studies, plus the two randomized experiments, in his "best evidence synthesis." Slavin found the mean effects of homogeneous grouping to be near zero for the 18 studies. Figure 2 displays each study's effect size estimate.

Figure 2—Effect Size Estimates of Middle School Math Studies That Compared Homogeneous and Heterogeneous Grouping



SOURCE: Slavin, R. 1993. "Ability Grouping in the Middle Grades: Achievement Effects and Alternatives" *Elementary School Journal* 93 (5), pp. 535–552.

But Do the Findings Generalize?

Some researchers (e.g., Elmore 1993) have questioned the potential of evidence from a "best evidence synthesis" to inform practice. Implicit in this question is the notion that a "best-evidence synthesis" (or meta-analysis) contains insufficient evidence to generalize to other settings. This notion is not entirely accurate. Slavin, for instance, includes one analysis (Hoffer 1992) that made use of data from the Longitudinal Study of American Youth (LSAY), a 4-year, large-scale study. He also says that other such studies "provide important additional information not obtainable from the typically smaller and shorter experimental studies" (Slavin 1993, p. 539). However, Slavin does not discuss the premise underlying cross-design synthesis, namely, that evidence from experimental studies and observational studies might be combined to generate more national estimates of effect. LSAY, however, may not be the ideal study for this purpose insofar as ability grouping is concerned. Adequate data, for example, were available from only 1,800 8th grade math students. NAEP's Trial

State Assessment, in comparison, collected data from about 2,500 8th grade students from *each* state.

What Does NAEP Reveal About Ability Grouping?

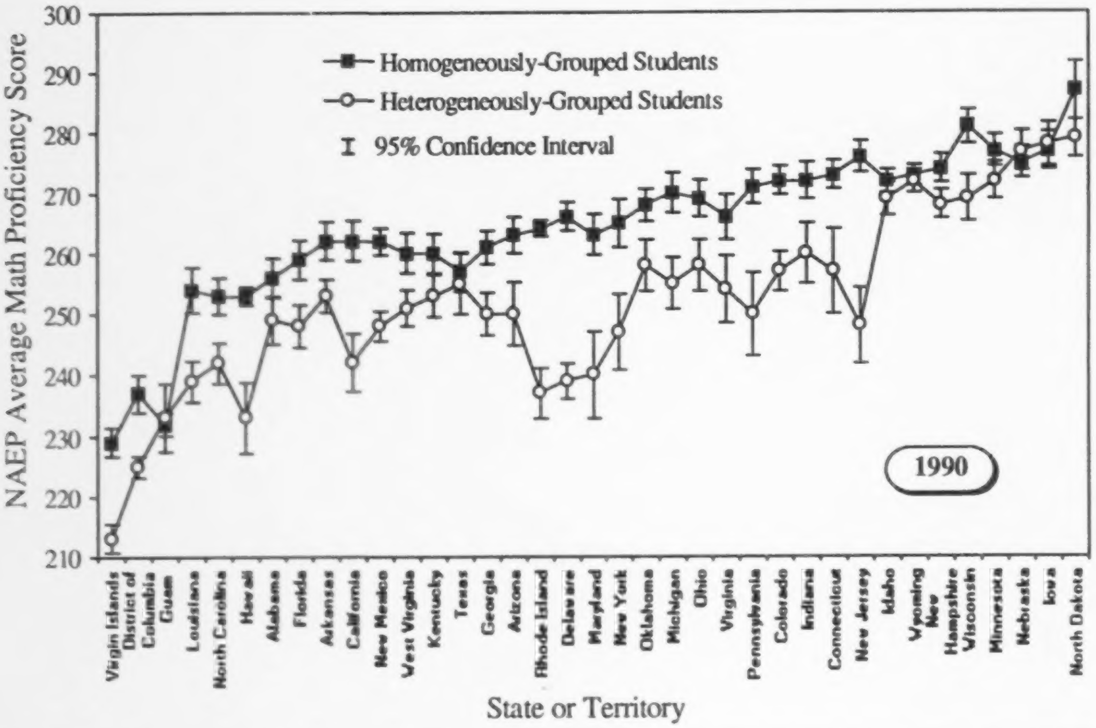
NAEP provides information on student achievement to local, state, and federal policymakers on a biennial basis. It also provides background information on students, teachers, and school administrators. Some NAEP information is demographic, while other information concerns educational practices and policies. NAEP allows policymakers to know, for example, whether student achievement is related to ability grouping. NAEP is an observational study, however. Making statements of impact or effectiveness on the basis of NAEP data is therefore inappropriate without some adjustment. It is imprudent to assume, for example, that students who are grouped by ability are comparable in all ways to those who are not. Schools, for example, may tend to group higher achieving students by homogeneous ability rather than heterogeneous ability, thereby causing an imbalance between groups that may bias achievement-based comparisons. *Unadjusted* NAEP data indicate, for example, that homogeneously grouped 8th grade (public school) math students outperformed their heterogeneously grouped counterparts in 34 of 37 jurisdictions (significantly in 27 of 37) in 1990, in 43 of 44 jurisdictions (significantly in 34 of 44) in 1992, and nationally during both testing years, as Figures 3, 4, and 5 show.

The Need to Adjust NAEP

If one is to use NAEP data to estimate the effects of ability grouping, then one must first employ a substitute for the randomization of controlled experiments, i.e., to assure that the groups do not differ systematically. The focus here is on a "propensity score" adjustment—a technique to produce analyses that approximate the results of a controlled experiment. As applied to the example of ability grouping, the analyst's first task would be to develop a statistical model—on the basis of theory, following the lead of others (e.g., see Hoffer 1992), possibly through stepwise logistic regression, or through some combination of the three—to compute each student's probability of being grouped by ability (homogeneously); that is, to compute each student's propensity score. This approach may benefit from recent advances in the statistical theory for estimating multilevel models. Version 4 of Bryk and Raudenbush's hierarchical linear modeling software, for example, will enable analysts to model categorical dependent variables while taking into account the multilevel nature of NCES data.

After deriving propensity scores, the analyst's next task would be to divide the entire sample into quintiles on the basis of these scores; that is, to subclassify students on the basis of their propensity scores.[3] The analyst could then compare the achievement levels of subclassed ability-grouped (homogeneous) and non-ability-grouped (heterogeneous) students. In a sense, this procedure would generate five estimates of the effect of homogeneous grouping. An example of one possible interpretation is as follows: *With respect to students who were most likely to be grouped by ability, no difference in achievement exists between those who were actually grouped by ability (homogeneously) and those who were not.* The analyst could then combine the estimates by taking

Figure 3—Estimates of math proficiency for grade 8 students by type of instructional grouping (i.e., homogeneous or heterogeneous) for each state or territory in 1990



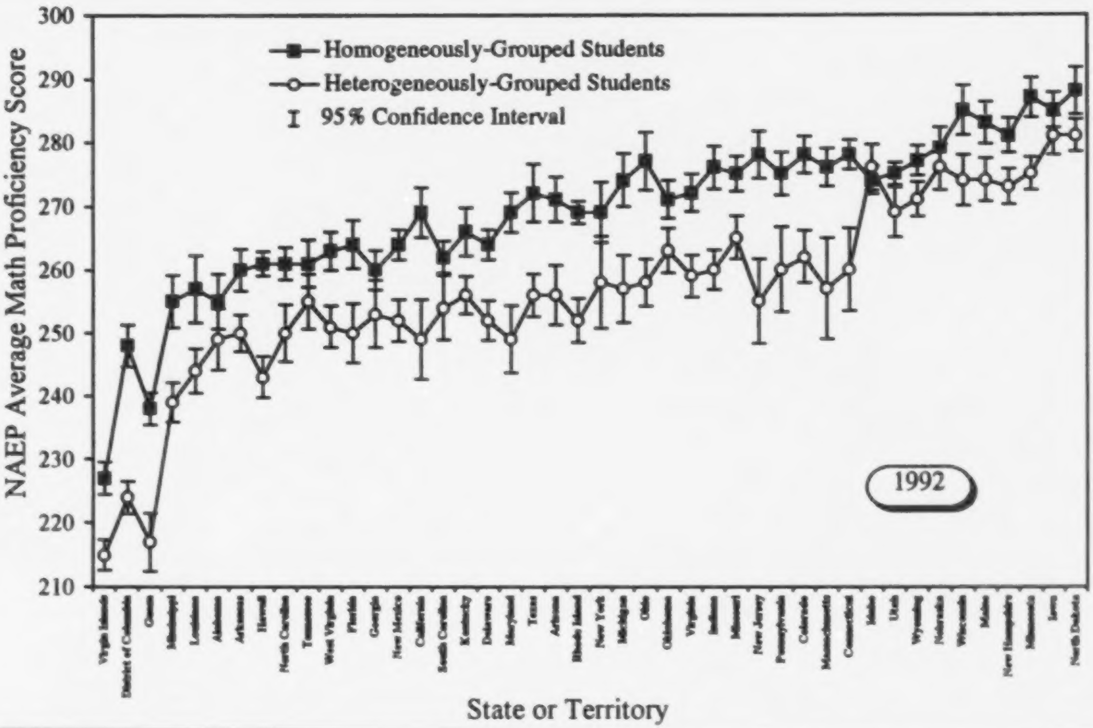
SOURCE: U.S. Department of Education, National Center for Education Statistics, 1993. *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States*, p. 463.

the average of the five effects, as in meta-analysis. This final estimate would be far more trustworthy than any of those that were displayed in Figures 3, 4, and 5.

Combining Contradictory Evidence: A Potential Problem

How one might *combine* estimates of effect from experiments and one or more NAEP analyses, particularly when the estimates are contradictory, is unclear. Although the GAO's introduction (1992) to cross-design synthesis discusses the problem and presents several options, it concludes that "many refinements are still to be developed" (GAO 1992, p. 96). The lone illustration (GAO 1995) of a cross-design synthesis, however, does not attempt to develop these refinements. The meta-analytic literature, which merited considerable consideration in the GAO's introduction (1992), also merits consideration here.

Figure 4—Estimates of math proficiency for grade 8 students by type of instructional Grouping (i.e., homogeneous or heterogeneous) for each state or territory in 1992



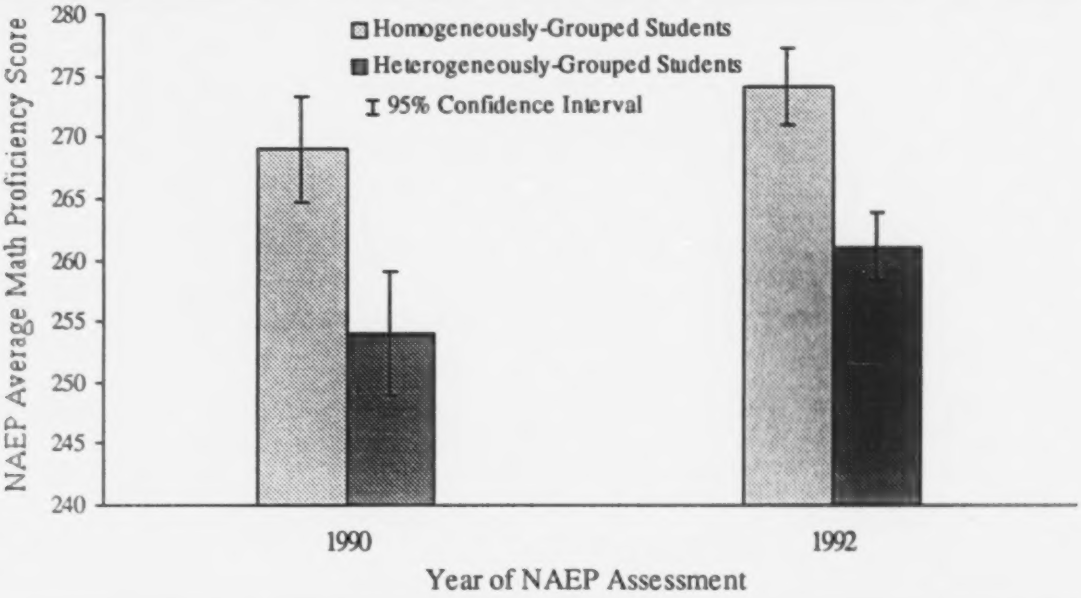
SOURCE: U.S. Department of Education, National Center for Education Statistic, 1993. *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States*, p. 463.

Meta-Analytic Strategy

In meta-analysis (see Hedges and Olkin 1985; Hunter and Schmidt 1990), the analyst computes one or more overall estimates of effect after 1) collecting, 2) coding, and then 3) weighting each study's effect size by its sample size; that is, the analyst computes a weighted average. This weighting scheme poses an analytic problem in the ability grouping example, however, on account of the size and nature of the sample of studies available for analysis. Put into question form: Does each adjusted NAEP state sample (of about 2,500 students) deserve to be weighted by 30 or so more times more than the smallest (Ford 1974, n=82) experimental study? The answer is probably not.

One approach to the problem is to divide the entire set of studies by design category (i.e., observational, randomized) prior to weighting studies within each category. After doing so, it then seems sensible to follow Hedges' and Olkin's advice. The general strategy that they recommend, as applied here, would be to do separate tests of homogeneity for the two sets—1) 44 state-level

Figure 5—National estimates of math proficiency for grade 8 students by type of instructional grouping (i.e., homogeneous or heterogeneous) in 1990 and 1992



SOURCE: U.S. Department of Education, National Center for Education Statistics. *NAEP Data on Disk: 1992 Almanac Viewer.*

NAEP analyses for 1992, and 2) two randomized experiments—of effect sizes. If, for the NAEP analyses, the null hypothesis of no difference is rejected (i.e., if significant random variation exists among the 44 effect size estimates), the analyst might then include additional covariates in the model to attempt to explain the variation. States with low teacher-student ratios, for example, may produce a small positive effect for ability grouping while those with high ratios may produce a large, negative effect. Teacher-student ratio (e.g., low or high) may therefore account for the variation beyond that expected from sampling error alone among all state-level effect sizes. The analysis, then, would generate two indices of effect for the NAEP analyses. Combining the two, however, would be inappropriate. The analysis would also produce (at least hypothetically because there are only two randomized studies) one or more estimates of effect for the experimental studies. This estimate or these estimates will be distinct from that, or those, produced by the experimental studies.

In this framework, it would also be possible to include in the analysis additional observational studies. Combining Hoffer's findings (1992) on the comparative effectiveness of homogeneous and heterogeneous grouping with those from the potential NAEP analyses, for instance, is one possibility. A reanalysis of Hoffer's data may be in order, however. Although Hoffer makes use of propensity scores, he does not use them to directly compare homogeneous and heterogeneous groups.

To summarize, *it may not be possible* to combine estimates of effect from experiments and one or more state-level NAEP analyses when the estimates contradict one another, particularly when there are very few experimental studies available for synthesis. As the sample of available experimental studies increases, however, the possibility of combining estimates across design categories also increases.

Implications of Research on Grouping

There are two broad implications of this illustration. First, there is an obvious need for more randomized experiments. Second, we will never know whether the apparent performance difference between homogeneous and heterogeneous groups is real without deeper analysis. It seems important to carry out the analysis, however, because the percentage of 8th grade math students who were grouped by ability decreased in 30 of 36 states (that participated in both NAEP Trial State Assessments) between 1990 and 1992 (USDE 1993)—a decline that may or may not prove wise, depending on the outcome of the proposed analysis.

HIERARCHICAL MODELS, MODELS MORE GENERALLY, AND THEORY: IMPLICATIONS FOR DESIGN OF NCES SURVEYS

Background

Survey samples sponsored by the NCES have often obtained data on institutions, such as schools, and simultaneously obtained data on individuals within the same schools, such as students. These include the National Longitudinal Study of the High School Class of 1972 (NLS-72), the National Education Longitudinal Study of 8th graders in 1988 (NELS:88), and High School and Beyond (HS&B), which focused on the high school class of 1980 and emulated parts of NLS-72.

The data on institutions have been combined in analysis with the data on individuals at times. Coleman, Hoffer, and Kilgore (1982) did do so, partly in the interest of discovering the relative effects of public versus private schools on student performance. Mosteller and Moynihan (1972) did so to understand the effectiveness of compensatory education programs. These illustrate early attempts to recognize the hierarchical nature of the data. More recent examples are not hard to identify, although Draper (1995) suggests that such analyses are the exception rather than the rule.

Despite the burst of recent attention to hierarchical data, technical advances in their analysis have been made for over 40 years (Draper et al. 1992). Work on the software that executes the analyses has been especially inventive and industrious over the last few years (Bryk et al. 1989). The fact that National Center for Educational Statistics has been collecting multilevel data for over 2 decades suggests that NCES anticipated, rather than lagged, advances in the software and analysis of such data, at least incrementally.

Draper (1995) argues that recent developments in hierarchical models (HM) have three clear advantages over earlier approaches to the statistical analysis of multilevel data:

1) "a natural environment within which to express and compare theories about structural relationships . . .

2) better calibrated uncertainty assessments in the presence of positive intracluster correlations . . .

3) an explicit framework in order to combine information across units . . . to produce accurate . . . predictions of observable outcomes."

Some readers are doubtless aware that Benefit #1 has been claimed for other analytic methods, such as LISREL. Boot strapping independent of HM arguably helps foster Virtue #2.

Hierarchical Scenarios, HM Models, and Analysis

The HM model we define here as a stochastic one that represents a setting in which units at the lowest level of measurement, for example, "A," are nested within units measured at a higher level, called "B," and these in turn may be nested in a still higher level of measurement unit called "C," and so on. Sampling and other random error is recognized at each level in the model. A variety of models and associated analyses can be regarded as special cases of a general hierarchical model.

So, for example, students (A) may be nested within classrooms (B) and classrooms may then be nested within schools (C). Variations among students, among classrooms, and among schools may be recognized in the random error terms and in other features of the model. Models that represent this scenario and the analyses are described in Bryk et al. (1989). An application to the data generated by the National Assessment of Educational Progress (NAEP) is given in Mullis, Jenkins, and Johnson (1994).

Or, time points of measurement (A) may be nested within students (B) who themselves are nested within classrooms. Some random coefficients models/analyses for longitudinal data fit this scenario. A related application to a sizable longitudinal study of participants in Boy's Town is given in Osgood and Smith (1995).

Or, one may conceive of a set of independent studies as a scenario in which individuals in (A) are nested within a given study (B) and various studies may be nested within (say) multiple geographic regions or institutions (level C). This scenario is similar to those encountered in attempts to combine evidence from different sources. Such a combination falls under the rubric of meta-analysis (Draper et al. 1992).

These scenarios and the associated analyses are considered in what follows. The emphasis here, as elsewhere in this report, is on what the advances in HM analyses imply for improving design of NCES surveys. The implications may concern: what units ought to be measured and how many, how, when and with what frequency, and at what level in a hierarchical setting (see Exhibit 1).

Hierarchical Models and Cross-Sectional Surveys

In principle, advances in hierarchical models (HM) invite one to analyze observations in contexts, e.g., students within classrooms, within school, within school districts, within states, and so on. An obvious abstract implication of the availability of the HM technology is that NCES might then collect data at these various levels. This data collection would be in the interest of exploiting a technology that purports to help understand, for example, how students' academic performance is influenced by classroom teachers, their schools, and the state education policies that influence them. At least, one might exploit the HM technology to understand where interesting statistical associations appear, even if one cannot be confident of where and how the influences are exercised, and even if one ignores time as a variable.

The implication just given is embarrassingly vague. It is also important. To get beyond the vagueness, we need to get to specific data sets, and to understand features of the models and the associated analyses and the data. Mullis, Jenkins, and Johnson (1994) did so. They tried out HM-based approaches to analyzing NAEP data on mathematics achievement. Their object was to identify "unusually effective schools" (outliers) and to determine how and why such schools differed from others. The bases for understanding were HM analyses that helped to arrange data at the student level within school and at the school level, so as to identify the predictable influences on student performance and school performance. Schools that departed from prediction in a positive direction could be regarded as unusually effective.

Here, the concern lies not with the substantive results of the Mullis et al. (1994) paper, which are interesting. Rather, the concern lies in what the authors say about better design of NCES surveys. The Mullis, Jenkins, and Johnson monograph, as one might expect from other sections of this report, contains no section on "implications." Drawing implications for better design of NAEP was not identified as an objective in the monograph.

Mullis, Jenkins, and Johnson did, however, construct a section entitled "Technical Issues in the Application of HLM to NAEP Data" (pp. 103-112). It is a springboard to implications. Their section taught us another small lesson: implicit in scholarly discussions of "issues" are possible implications. It invites us to encourage authors to write about issues rather than implications.

In short, what does the "technical issues" section of the conscientious HM analysis by Mullis, Jenkins, and Johnson imply for better design of NCES surveys?

What Should Be Measured: Implications

First, NAEP measures of socioeconomic status (SES) are imperfect. Mullis et al. (1994) used what they could in a HM analysis based on NAEP. The imperfection in measuring SES are greater in NAEP than in other surveys. The implications are that

1) NAEP might measure SES more directly, e.g., asking questions about family size and income; or

2) NAEP might be linked to other information that gets at SES information more directly, e.g., SSA, IRS, and so on; or

3) NAEP might exploit imputation methods and/or indirect estimators to produce the SES information on individual students.

None of these options may be feasible for NCES. Still other options, not identified, may be more feasible. The raw implication is that analyses of NAEP data would be better if data on SES in the NAEP samples were better.

Who and How Many of Them: Implications

The Mullis, Jenkins, and Johnson (1994) report says plainly that the number of teachers within schools was not sufficient to sustain a HM analysis that could recognize the influence of classrooms and teachers (p. 104, first full paragraph). The implication here, as elsewhere, is conditional. If NCES and its clients want to learn about how teachers (classrooms) influence student behavior, having taken into account student-level variables such as family backgrounds and school-level variables, then NAEP should be designed so as to get at this level. That is, more teachers per school should be surveyed where multiple teachers per grade or class is the form.

The Mullis, Jenkins, and Johnson (1994) report also recognizes that the number of students within each school in NAEP may not be sufficient to estimate within school parameters (p. 104, last paragraph). Roughly speaking, they recognize, as others do, that relying on a random sample of 15-20 students within a school may not be sensible if the object is to understand average 8th grade students' performance within the school. But they also recognize that these data and estimates of average performance are aggregated up to regional and national levels that are arguably reliable because there are so many schools in the NAEP sample—1,500 schools in the aggregate.

We are aware of only one study of sample-size design based on HM that may be worth building on, by Magdalena Mok (1995). She chose a simulation scenario that is concrete, but it may not accord with scenarios in North America. Mok's simulation approach is at least promising, despite debatable relevance of the particular scenario.

This matter of numbers is controversial. The cautious implication is that NCES should support an investigation of sample size at all levels in the HM context. There appear to have been no comprehensive studies of statistical power/sample size issues or at least none sufficient to inform adjudicate decisions at the design stage of NAEP.

Longitudinal Data Analysis

Studies in education often explore how entities change over time. Analysis of such data has improved on account of analytical statistical advances in understanding growth curves, random coefficient models, event history, and so on. Analysis of longitudinal data on individuals is a special case.

Rogosa and Saner (1995) clarified approaches to analyzing such data and have compared different analytic methods. Breslow's paper (1989), unlike that of Rogosa and Saner (1995), stresses the benefit of empirical Bayes estimates over OLS in the context of longitudinal study in biometry. We depend here on the Rogosa work to lay out crude implications of the approach for the design and improvement of NCES surveys.

The presumption is that understanding individual growth is important, inasmuch as questions about growth precede and drive the exploitation of random coefficients (or other models) in analyzing the data. If NCES professionals, educational researchers, or other users of NCES data declare that questions about growth are unimportant, then the implications drawn here are unimportant.

The rudimentary individual growth model posits that the individual's state at time *t* is a simple linear function of time and random error. The individual's outcome state may be measured with error. It is common to characterize measures of this outcome using a classical measurement error model. Each individual in a group is characterized by the individual's base intercept and his or her growth parameter, i.e., a linear regression of outcomes on time. The model that describes this also recognizes random error. The group of individuals is then characterized by an overall mean and a mean growth parameter and some index of variability within the group over time.

This basic model is augmented, at times, by assuming that the individual's growth is a function of certain other variables. The individual's participation in a compensatory education program or the hours that the student spent studying are illustrations of such "control" variables.

Crude Implications

NCES should figure out *when* to measure each individual. Measure each individual's state at each of the time points, e.g., achievement, record each time point, and measure exogenous variables *z* that may influence growth parameters. These broad implications are obvious. Rogosa and Saner (1995) and others raised questions that bear on more interesting implications of analytic work on understanding growth.

Less Crude Implications: Sample Size

Empirical and simulation studies suggest that small samples lead to intolerably large standard errors in estimating growth parameters. Sample sizes above 200 seem acceptable to Rogosa and Saner, given the kinds of questions that they have explored. At the national and state level, NCES routinely depends on larger samples.

How big should the sample size be, under what conditions and particular growth models, and with what particular method of estimation? As yet, there seems to be no general answer to the question. This question can have no specific answer absent a specific question about what needs to be understood about a specific phenomenon. NCES may then choose to wait for others to address this question before going further. It may sponsor special studies to address the question so as to serve contemporary interest in growth curve analysis.

Less Crude Implications: How Often to Measure

Rogosa and Saner (1995) suggested that 4-6 time points for measurement is not uncommon. But we have seen no substantial analytic, empirical, or theoretical handling of the topic of how an agency such as NCES should decide.

If education would be served well by research on individual growth curves, then the study of when, how often, why, and how observations should be made is sensible. NCES might then commission studies that lay out the issues and support pilot work that addresses them. Or, NCES may wait for others to proceed further.

Hierarchical Models and Meta-Analysis

Draper (1995) considered briefly the link between hierarchical models and meta-analysis. He cited the hierarchical model's ability to detect between-study variation as the main reason why it is "a natural tool for implementing [a meta-analysis]" (p. 133). In the discussion, Draper described a six-study analysis (Goodman 1989) of the effect of aspirin on the survival rate of patients who had survived a heart attack. Although the results of the meta-analysis suggest that treatment is effective, there was substantial between-study variation. The researcher who initially implemented the meta-analysis, however, did not then "pose and [test] a series of linear models to explain the variation" (Bryk and Raudenbush 1992, p. 156).

From Draper's perspective, this constitutes a misuse of the hierarchical model. He contends, that "this can actually promote an antiscientific attitude of indifference to the cause of the study-level discrepancies" (p. 134). The implications of Draper's perspective "for allocation of research effort and resources" (p. 133) are to invest research time and money in discovering how and why the study level characteristics explain the between-study variation *before* recommending treatment. The implications of Draper's perspective on meta-analysis for the design of NCES surveys, however, are for the most part less clear.

Modeling and Analysis Generally

Clogg (1989) identified points of uncertainty in constructing models in the social and behavioral sciences and education. Each point engenders difficult choices in analysis. Each choice might be better informed through better survey design. Freedman (1985) assaulted conventional approaches to modeling in the social sciences, including those in education. Freedman's scientifically assaultive approach and Clogg's empirical approach have some of the same implications.

Universe

Clogg maintained that data analysts who depend on survey data produced by statistical agencies need better information about the universe that is sampled than they usually have. Because so much analysis is directed toward making generalizations about the nation based on national probability samples, he argues that the *census* must be improved. For instance, the census often is used as a benchmark for checking the quality of other surveys, including NCES surveys. If the census universe is imperfectly specified, then the benchmark checks will be misleading. Similarly, if census figures are used to construct sampling frames but certain groups are undercounted, then the frames will produce results that differ from what they should be.

The implication is that insofar as NCES relies on census figures to design its surveys, improvements in the decennial census can help to improve NCES survey design and the analysis of NCES data.

Measurement

Clogg (1989) believed that good measurement is fundamental to good analysis and praised contemporary cognitive research on asking questions. Freedman (1995) argued that good measurement is not common enough in the social sciences. He told us that "good models are hard to build on bad data." But, aside from criticism of factor analysis, Freedman told us nothing new.

Cognitive approaches to understanding how people respond to questions can be regarded as a new approach to analysis and to designing better surveys. They may be employed at the survey design stage and, indeed, NCES does so. There is little published on the product of the effort however. The approach might also be productively employed at the stage of statistical analysis. For example, such research might reveal why nonresponse rate is relatively high for teachers' responses to questions about credentials in the Schools and Staffing Survey; these may then lead to redesign of the questions.

There are at least two implications with regard to cognitive research. First, publishing on the lessons learned from earlier NCES investments in predesign work on cognitive aspects of questions seems sensible. This might be done through NCES *Research and Development* reports or other means. The product is arguably of potential value for all scholars who seek to pattern local surveys after NCES efforts. The second implication is that the cognitive research might be undertaken after the survey is done in the interest of better understanding of the survey's results. NCES might encourage this at low cost to the agency through a variety of means—predoctoral and postdoctoral fellowship work, collaboration with university-based or institutional researchers, reliance on able and thoughtful graduate students, and so on.

Complex Sample Design

For Clogg (1989), "the failure to take account of uncertainty produced by complex sampling procedures is surely one of the most embarrassing problems we have at the moment. For at least some cases, reasonably tractable procedures are available, but the technology available now seems difficult to implement in the context of the formal models that we estimate routinely."

In some respects, NCES has already invested productively in addressing Clogg's concern. Scholars who seek to analyze data from the Schools and Staffing Survey (SASS), for example, are supported in effect by software (based on still other analyses) that characterize uncertainty in estimates of parameters *and* in formal statistical tests of a conventional variety.

The implication is that NCES ought to continue to build more user-friendly and accurate characterizations of uncertainty.

The Interface Between Theory and Models

For Clogg (1989, pp. 218-19), "the goal of analyzing social statistics is to explain how a system of variables works." The idea that theory is important is implicit in his remarks. Zellner (1989, p. 164), in discussing successful modeling of the sort that Clogg describes, says: "a good deal depends on whether good, relevant statistical theory and subject matter theory are available." That is, without good subject matter theory, modelers are forced to be content with description and exploratory work that may help to illuminate the structure underlying data and to make forecasts.

The immediate implication is that where subject matter theory is good, new analysis methods and models that generate the methods can be used to explore the theory. The products of this activity may have implications for better surveys.

The broader implication is that NCES should be aware of theories for which new analysis methods are useful. This awareness might be achieved, as it is at times, through advisory groups and consultants. It generally is achieved through contractors only when the contractors contribute to theory and to responses to an RFP (Coleman).

Subject matter theory in education in some areas is not sufficiently specific to determine which models ought to be used. This forces us to think in terms of description and forecasting. The less obvious and less certain implication is that designing better surveys rests heavily on deciding what to describe and how to describe it, rather than on new analytic methods and models.

The Roles of Models

Suppose we consider surveys of the kind that NCES runs as "nonexperimental social science." Suppose we then consider "new models" and the analyses they engender and ask: What the role of such models is in nonexperimental social science?" In fact, the question has been posed and addressed, in a special issue of the *Journal of Educational Statistics* (Shaffer 1992). The primary new models and methods reviewed in Shaffer (1992) include path analysis and structural models. David Freedman provided the main criticism. Rejoinders and reactions were developed by among others, David Rogosa, who is also not well disposed toward such models, and by Peter Bentler, Herman Wold, and others who have tried to develop such models.

Direct Implications

The volume contains no direct discussion of how structural models, as represented by LISREL or EQS, for instance, or of path models, should influence the design of observational studies.

Indirect and Very General Implications

The advocates of structural models argue that they are useful in developing parsimonious description. In effect, one is able to characterize a measure on any array of variables as unobserved measures on a far fewer number of latent variables. Then one could construe an implication as we should assure that the number of variables is sufficient to identify the latent trait well. If one believes that "home environment" is important as an "unobservable trait" of a child, there must be a sufficient number of questionnaire items to get at it.

Conversely, one may have many questionnaire items and reduce their number rationally through some approach related to structural models, e.g., factor analysis. How well one might do this depends heavily on substantive theory about how latent variables are related among themselves, and to the variables (questionnaire items) actually used. Freedman argues that good theory is absent. Further, LISREL and related approaches are not theory construction methods. He and Rogosa argue further that the scientific approaches are questionable at best.

More Direct Implications That Are Negative

Rogosa argued that if understanding individual growth or change is a main objective, then models/methods such as path analysis and structural models are inappropriate. His "message is that the between-wave covariance matrix provides little information about change or growth" (p. 89). More to the point for science, "covariance matrices arising from very different collections of growth curves can be indistinguishable" (p. 93).

The longitudinal studies of NCES are developed, in part, to understand individual change. Rogosa's position suggests that because structural models are inappropriate for analysis of data from observational studies, they are also an inappropriate resource for guiding the design of observational studies. Such methods have no implications for design because they are irrelevant to sensible analysis.

Rogosa argued further that more transparent and defensible approaches to growth curve analysis are at hand. In particular, the good scientist and statistical analyst can model each individual's trajectories, estimating parameters within each individual. One then models the differences among individuals.

Implication: Causal Structural and Path Models and Methods

Over the last decade, NCES has been advised not to undertake analyses that are causal in their orientation. The advice has been rendered by one Advisory Council on Education Statistics, notably by ACES Chairman Ellis Page during the 1980s. There was active opposition to such analyses by USDE Undersecretary Chester Finn to NCES' conducting such work in the late 1980s.

The models considered here and related ones are regarded as important in some quarters, e.g., among some education and sociological researchers. They are regarded as valueless, absent stronger theory, by some educational and sociological researchers. The controversy is sufficient to justify educing that, exploiting the models and methods ought to be avoided, at least in official reports on the state of education.

A second implication bears on advisory committees' appointee to provide counsel on surveys. In particular, NCES may exclude from its survey advisory groups' scholars whose special interests lie in building structural models, path models, etc. on account of the controversiality of the latter. Or, such individuals might be included provided the advisory group is augmented by those who believe that such models are useless. Individuals with an interest in structural models arguably enhance the likelihood that variables regarded as important will be collected in an NCES survey. The inclusion of opponents will temper that influence.

The exclusion of structuralist fosters counsel based on a perspective that is more prescriptive, e.g., how children grow or change, or how districts grow and change and what are the covariates of growth. The models and methods then are arguably more transparent. The implications may be more obvious, e.g., focus on collecting data at more time points rather than fewer points and more subnational data.

The implication then is complicated. To the extent that NCES regards its role as the production of informative descriptive statistics (that are exploited occasionally for causal analyses), than relying on models, analyses, etc. that are not causal in their orientation is important. We can understand a lot about growth and change by seriously observing growth and change, not only through questionnaire/telephone surveys and administrative records, but through more direct observation.

To the extent that NCES regards its role as fostering the opportunity for structural, path, causal, analytical, *and* descriptive statistics, then relying on the more complex models is sensible. Society is complex and models must presumably be more complex.

There may, in fact, be no real conflict between these options in at least one sense. The data produced on the basis of an orientation toward good description (e.g., simple growth analysis) may "satisfice" for the structural models who use NCES data. Learning whether there is a satisfice and whether there are major differences in what each group regards as satisfactory suggests that NCES bring the groups together.

Implication: Vernacular and Causal Models

The Human Genome Project has had the benefit of great talent and considerable resources. Despite this, what a gene is, what "genome" means, and what the adjective "genetic" implies are still subject to some debate (*Science* 1994, 1995). Just as the scientists and statisticians struggle with vernacular differences and with remarkable efforts to understand what one means in the genetic arena, NCES and others must confront ambiguities in the model-building arena.

It does not seem unreasonable for NCES or scholars outside NCES to be confused about some statistical models and analytic methods. For example, "structural models" have been defined at times as models whose parameters are invariant across some space, e.g., over time or geographic area. Structural models have *also* been defined, more or less, as models that represent the "as if by experiment." That is, they have come to be regarded as causal models.

Historically, "path models" have been characterized as "causal models." This characterization is important. These models, and structural models that are conceived of as causal models, are ways to develop a story. The story is one of plausible explanation of what influences what. Some scholars, such as Rogosa, have grouped the path models with structural models at times. Path and structural models have been lumped in with analytic methods whose underlying models were not originally explicit, e.g., cross-legged panel analysis.

At least a few scholars have tried to make distinctions plainer. Freedman distinguished between models that are helpful in summarizing data and those models that are born of more ambitious objectives, e.g., structural models. Roughly speaking, the idea is that observations on some "X"s are empirically related to observations on some "Y"s. Further and more important, Freedman's argument is that this is good description, but does not necessarily meet standards for a good structural model. The structural model is one that represents a good scientific theory.

The main point is that very able people, people who think, suffer the consequences of vernacular. "Structural models" for a fine economist may/can mean something different for a fine statistician or psychologists. "Causal models" means something to those of us who try to encourage controlled randomized experiments. It means the same, but it also means something different, to those of us who try to understand what variables (X) influence what variables (Y) in what is theory-based and arrangement based on observational data.

Exploiting Theory That Drives Survey Design and Model-Based Analyses

New analytic methods in statistics yield few *specific* implications for designing surveys. This is despite their usefulness for *interpreting* the data collected on the basis of an explicit design. Why is this? One can argue that we should not expect new analytic methods to imply anything about designing research. After all, the new methods have a certain objective, e.g., developing an estimate of a parameter that is better than its competitors. The first object is generally not to produce a better design.

Suppose we take a step back and imagine that advances in substantive theory (or policy), rather than the new analytic methods, are the drivers for improving both survey design and the analytic methods. Consider a simple example. Gender, in theory, is important. This theory then drives design of surveys that permit one to say something about gender differences. The data from such a survey permits one to analyze gender differences and to perhaps improve analysis.

This line of thinking may seem obviously true, at least, to a theoretician in the education arena. It may not be obvious to others. Even if the line of thinking seems sensible, how do we exploit this in the interest of better survey design? "Theories" are in ample supply in the education arena. Merely saying that we ought to rely more on theory to advance survey design is gratuitous.

It seems sensible to ask two questions: "Has a theory-based approach helped us to learn about how to improve design? If so, how do we exploit theory better, given the abundant supply?"

Consider one example by way of addressing the first question: How has theory-based analysis helped? Boe and his colleagues have depended on the NCES Schools and Staffing Survey (SASS) to understand the relation between teacher supply and demand, and the flow processes that underlie the relationship. Their analyses depend partly on being able to enumerate, from NCES data, teachers who are not credentialed to teach, i.e., anybody who is part of the actual supply of teachers. Their analyses cannot recognize a special source that is arguably important on theoretical and policy grounds. This source includes individuals who have been trained and employed in one professional arena but who then move to another. Engineers and scientists involved in the defense industry, for example, have moved to other occupations on account of the reduction in size of the industry. Some of these people find their way into the supply of teachers through federal programs that foster their transition. That is, thinking (theorizing) and finding out based on the thinking has an implication for SASS: the survey ought to obtain information about how certain people find their way into the teacher supply.

Suppose the reader is willing to grant that theory should have some influence on survey design and the models and analyses that exploit the resulting data in turn improve theory, design, and so forth. How might NCES be instrumental in tracking advances in theory so as to facilitate advances in design?

Implication #1

First, some framework for understanding advances can be invented. A simple one, based on the generic list given earlier for NCES' tracking advances in analytic methods (Exhibit 1), might simply list the things that new theory can address:

- What (new) variable ought to be measured?
- How and at what level ought the variable be measured?
- Who (or what entity) should be measured?
- When and with what frequency and periodicity should a variable be measured?
- What stratum (kinds of individuals, entities) should be observed?
- What statistical relationship need to be examined?

For example, some theorists have argued that we need to know more about family environment to understand the nature of family and school influences on the children's education. This implies that NCES can measure more related variables or measure them better in some NCES surveys that lend themselves to analysis of the topic. The NCES longitudinal surveys are an obvious option.

Considering the matter of statistical relationships, some analysts of NCES data have argued on scientific grounds that survey variables that are unrelated to others are candidates for abandonment. The argument is plausible.

In each of these cases, and others that can easily be constructed, theory plays a role. And a simple framework for understanding progress in thinking seems important for NCES and perhaps its sibling organizations. The mechanisms for tracking incremental advances on each front are fragmented.

Implication #2

Tracking advances in theory demands that NCES choose a target. Identifying self-declared theorists would result in a population of informants, of course. If these scholars depend directly or (more likely) indirectly on analyses of NCES data, then such an effort might be productive for NCES in the short or interim term. It then seems sensible to be able to locate and make use of individuals who actively capitalize on NCES data *and* individuals who depend on these data. The basic mechanisms available to NCES for doing so include those that NCES already depends on: using members of advisory groups for specific surveys, for projects undertaken by NCES contractors or by the NAS list of users of NCES data, and so on.

Implication #3

Assuming that relevant scholars can be identified and that frameworks for tracking the advancement in theory can be invented, then some method to facilitate the acquisition and sharing of information is still necessary. Conventional research journals, meetings of NCES data users and advisors are vehicles for doing so, that NCES already depends on. We might add to this the possibility that NCES can take better advantage of the World Wide Web. Such options are presented in the section on new technology.

COUNTING THE HARD TO COUNT AND MEASURING THE HARD TO MEASURE

Two topics are considered here. The first concerns eliciting information from respondents about a sensitive trait, state, or event. The second concerns the measurement of mathematics ability in NELS:88.

Background: Counting

At times, NCES has elicited information from students and others that can be regarded as sensitive. For example, NCES has asked students whether they have been victimized in an assault.

Asking a student member whether he or she provoked a fight or assaulted another student would be regarded as more sensitive. Asking about their having stolen property, engaged in unprotected sexual activity, and so forth may be regarded as extremely sensitive.

The Congress has attempted to limit the extent to which sensitive information can be required from students in surveys without the consent of their parents. Section 439(b) of the General Education Provisions Act (20 U.S.C. 1232g) for example, was amended in 1995 to say the following:

> No student shall be required . . . to submit to a survey, analysis, or evaluation that reveals information concerning . . . mental and psychological problems potentially embarrassing to the student or his family . . . sex, behavior and attitudes . . . illegal, anti-social, self incriminating behavior . . . income without prior consent of (adult or emancipated minor) or . . . of parent (minor).

That many surveys run by NCES are voluntary, rather than required of students, makes this statute a bit peculiar in its value. But recognize that the voluntariness of a survey may not be understood. And in any case, the mere posing of questions to a student about the student himself or herself may be offensive to some parents or teachers. Schools may decline to cooperate in surveys because the information being elicited from an individual about the individual's own behavior is regarded as sensitive.

Suppose that in many cases, NCES will not be able or willing to elicit sensitive information directly from students or others about their behavior. How then might one obtain information sufficient to estimate the incidence of a sensitive characteristic or behavior?

Network-Based Estimates

One approach to the problem of eliciting sensitive information has been developed by quantitative anthropologists and others with an interest in counting the hard to count: network-based estimators. Roughly speaking, individuals in a sample are asked *not* about their own behavior. Rather, they are asked about the behavior of unidentified acquaintances in their social network. For instance, we may ask a student: "How many students do you know provoked a fight in the last month?" This question is proffered instead of: "Did you provoke a fight during the last month?"

To estimate the total number of students who provoked fights, one also needs to elicit information from students about the size of students' social networks. The estimate of network size may be based on a survey question or a separate side study. Data on provoking fights elicited from students in the survey is combined with data on the average size of students' social network and on the size of the student population to produce an unbiased estimate of the total number of students who provoke fights.

Understanding how to estimate the average size of a personal network is no easy matter. It arguably depends on what kind of persons that one might encounter in a sample. For instance, a probability sample of adults might include priests and mail deliverers whose acquaintanceship network is larger than, e.g., a cloistered monk's. The efforts to understand personal network size in a variety of studies are reported in Bernard et al. (1987, 1989, 1990), Killworth et al. (1990), and others given in the reference list attached.

Contributions in this arena lie partly on the design side of surveys, e.g., learning how to elicit information. Part lies in analysis, including constructing estimates and understanding their quality.

Prior Analytic Work: Empirical Studies

The network-based approach has developed in research over the last decade or so. One of the more recent applications and a test of the approach is reported by Laumann, Gagnon, Michaels, Michael, and Coleman (1989). Their object was to estimate the prevalence of AIDS in the U.S. using the network approach. This was done partly in the interest of assessing the Centers for Disease Control's estimate of prevalence. The authors' vehicle for judging the quality of the network-based approach was a comparison of a network-based estimate of the distribution of homicide victims made against the distribution yielded by the FBI's Uniform Crime Reports and the PHS Vital Statistics. Network-based questions were embedded in the larger context of the 1988 General Social Survey, a survey that is independent of the FBI and PHS.

The results suggest that the network-based approach is trustworthy in producing homicide rates that are close to those yielded by official crime statistics. If one then chooses to trust the network-based estimates of AIDS prevalence, it appears then that the CDC data overestimate prevalence in some categories (e.g., in minority populations) and underestimate prevalence in others (e.g., in the Midwest).

The network-based approach has been the target for other interesting empirical research. For example, how to estimate average network size is crucial, and Bernard, Killsworth, and Johnsen (1994) have reviewed recent work.

A Broad Implication

The broad implication for NCES is this. If NCES wishes to estimate the number of people who have a sensitive characteristic in ways that avoid direct knowledge of the individual, then network-based estimators ought to be explored. In principle, the approach can be used in any NCES survey in which the size of the target population is known and a probability sample of the sample is drawn and the information about "knowing others who did X" and network size can be obtained. This then includes new waves of NELS:88, the National Household Education Survey, the birth cohort survey being considered, the Beginning Postsecondary Students Longitudinal Study, and at least some surveys mounted by the Fast Response Survey System.

Implication for the NCES National Household Education Survey

The National Household Education Survey presents some opportunities to exploit network-based estimators. NHES is based on a survey of a national probability sample of over 60,000 households. The sample and target population are well defined and over time (biennial roughly).

Assume that the target topic is sensitive, and that it would be difficult or impossible to get at the topic directly. Such topics might bear on the following:

- Indictment/conviction of school board members for wrongdoing;
- Indictment/conviction of teachers or staff who abuse children;
- Indictment/conviction of students;
- Parental abuse or neglect of students; and
- Who has been raped.

For instance, a survey of school board members that asks each individual whether he or she has been indicted or convicted of misuse of school funds, would arguably not be sensible. Obtaining information about the matter may nonetheless be desirable for *some* users of NCES data. Network-based estimators might be helpful to meet those users' demands. Moreover, they do it in a way that avoids privacy problems, embarrassment for the respondent, or intimidation. That is, asking each household respondent in a survey how many school board members they know have been convicted is likely to be more feasible than asking school board members whether they themselves have been convicted.

Asking the household respondent how many school board members or people they know also seems feasible. Similarly, understanding how many parents physically assault their children seems important. But the understanding cannot be gotten at directly. Instead, NCES might ask respondents in the Fast Response Survey or other vehicles whether the respondent *knows* about an assault, and about their network size (or independently, about network size in the target sample).

Background: Measuring the Hard to Measure

Kupermintz et al. (1995) analyzed the data from NCES' National Educational Longitudinal Study of 1988 (NELS:88) to understand how we might enhance the "validity and usefulness" of the NELS:88 measures of mathematics ability in the United States. Hamilton and colleagues (1995) examined the NELS:88 data on science also to understand how to enhance validity and reliability of testing on science.

The papers were generated as part of a seminar at Stanford. They are distinctive in that their objective was to enhance the quality of a periodic survey, notably NELS:88, based on conscientious analysis of data produced in earlier waves of the survey. Few published scholarly papers do so. There are still fewer that exploit "new analytic methods," regardless of how this phrase is defined, to do so.

Mathematics and Science in NELS:88

The work depended on a combination of small-scale cognitive research on the tests, conventional factor analysis, and new developments in full information-factor analysis. The latter involves employing a multidimensional item response model and a latent factor structure model that are, in conjunction, purported to yield estimates of item factor loadings on distinct abilities measured by the test that are better than estimates produced in other ways (Bock, Gibbons, and Muraki 1988;

Wilson, Wood, and Gibbons 1991). The implications of the latter are not obvious in the absence of its application to data such as mathematics tests in NELS:88. Even then, the relative contributions of specific analytic approaches used to yield the conclusions reached by the authors are unclear.

Broad Implications

What can NCES and we learn from this effort by our colleagues at Stanford? What are the implications of their work? The first lesson is that some university scholars indeed recognize that analyses can produce implications for better, not necessarily more, surveys. Further, and more important, they try to educe the implications. They are willing to embrace the challenge of doing so.

Second, the analyses are substantial, intriguing, and ecumenical. But the analyses occupy far more space in the published papers than do the papers' sections on conclusions. For readers who are interested in implications, this is not satisfying. The coverage is unbalanced, especially if the titles of the papers are taken seriously.

This perception of imbalance may be wrong, of course. An excellent implication based on ferociously difficult and time-consuming analyses, described in agonizing detail, may not take up much space. The idea that $E = MC^2$ is a conclusion of this sort. Despite considerable work, good conclusions that carry many implications can be astonishingly brief.

The analysts argued that the different dimensions of mathematical and science ability are influenced by different processes. Roughly speaking, the student's crystallized knowledge in mathematics is alleged to be influenced more by formal schooling; the fluid reasoning is influenced more by home factors. This argument is based on theory and on empirical regressions of mathematical factor scores on independent variables. The specific implication is that the theory and analyses ought to drive selection of variables and improvement of measures in NELS:88. For instance, one might focus more deeply on better measures of home education processes or characteristics that predict or explain fluid reasoning, keeping this initiative separate from attempts to develop measures of classroom processes that influence such reasoning. More generally, the implication that Kupermintz et al. (1995) draw is that large-scale assessments should certainly aim to represent the cognitive and educational distinctions being made by cognitive psychologists, math educators, and by the nation's education goals (p. 552).

That is, the theorists must be invited to contribute more to test development efforts and to the development of measures of potential influences on different abilities, especially higher order reasoning.

The Hamilton et al. (1995) results of analyzing science test scores from NELS:88 reveal different factor structures in 8th and 10th grades, and more factors than the math study revealed. The reasoning/knowledge distinction that appears clearly in analyzing the math scores does not appear in the science scores. Reasoning with knowledge appears as a factor distinct from science reasoning. The implication for the authors is that the multidimensional character of the tests ought to be recognized in reporting and in comparisons, e.g., by state. Further, the authors imply that the way the science abilities are measured ought to be augmented by cognitive studies of the way students respond to items; they believe these can inform test design.

Finally, Hamilton et al. (1995) reiterate the idea that design of the NELS:88 survey should be "linked" with more direct investigations of the context in which instruction take place. This may involve asking more questions about instructional practice, e.g., emphasis on discovery learning or reciprocal tutoring, or more likely, out-of-school activity that has theoretic and empirical relation to factor scores.

Even Broader Implications

Consider first the possibility of improving the mathematics assessments in NELS:88. Kupermintz et al. (1995) suggest that the current test is "multidimensional and should be treated as such" (p. 550). The bottom line is that they provided good evidence to suggest that the mathematics assessments currently in use get at both factual knowledge and reasoning or crystallized and fluid knowledge. Further, they argue that the finding ought to be taken seriously in improving new NCES surveys. Their implication is a little unclear:

> In general, future survey testing efforts should rethink intradomain distinctions among such achievement dimensions and their links to theoretical formulations and empirical findings on the structure of cognitive abilities (p. 550).

That is, NCES is not measuring one "thing"—mathematics ability. NCES measures several things that are tied up in mathematics ability. The recognition can come about through improved design of the tests, reporting, or in other ways.

A second broad implication for NCES survey design and for university-based education hinges on the way Dr. Richard Snow and his colleagues appear to have approached their task. The Stanford seminar focused on a specific data set, NELS:88, and a reasonably specific implications topic, improving math and science measurement in NELS:88. Further, funding was available through a competitive peer review grant awarded by OERI to sustain the effort. This strategy is not common but has a good pedigree; recall that the Moynihan-Mosteller work on equality of educational opportunity depended on a Harvard seminar series that engaged very able people. Figuring out how to do this right in the interest of NCES, students and professors, and the public is not easy. But the example is sufficiently instructive to encourage taking the time to think about the strategy, its value for NCES and the public, and more interestingly, its value in advancing science at large.

INDIRECT ESTIMATES, INCLUDING SMALL AREA ESTIMATORS

Background

Agencies such as the National Center for Education Statistics (NCES) obtain databases on national probability samples and generate statistics pertaining to the national level based on those data. For NCES and other statistical agencies, there has been episodic pressure to generate statistics at the subnational level, based on the national data. NCES also collects data at times at the subnational level, e.g., from the states. NCES has been encouraged at times to produce statistics at the substate level, once data users have found that the state-level data are instructive.

National samples, unless specially designed, do not usually yield results that are applicable to the state level. The national estimate of incidence of classroom disorder developed in NCES' Fast Response Survey of Public Schools on Safe, Disciplined, and Drug-Free Schools (1992) is not necessarily the incidence in New Jersey, for example. Similarly, the data on NAEP collected by NCES on students in Pennsylvania, for example, provides an estimate of mathematics ability for students in the state. The state-level estimate may not be an accurate characterization of abilities in local jurisdictions, such as Pittsburgh or Philadelphia.

The Question

Is it possible to exploit the data obtained at some aggregate-level data *and* other information so as to produce defensible estimates at the subaggregate level? Further, if it is possible, how can we understand the validity of these estimates?

Partial answers to the questions have been developed through recent work on Indirect Estimators. In what follows, we rely heavily on the Office of Management and Budget (OMB) (1993) *Statistical Policy Working Paper #21* and some other sources identified below.

The Approach: Indirect Estimators

Statistical agencies usually rely on *direct estimators* for reports. That is, the estimator, such as a mean number of assaults on students for the nation, is computed for a particular time and only from a sample of units in the population (or "domain") of primary interest, e.g., the students in the nation.

An *indirect estimator* is one that uses the design-based survey or a database for a direct estimator *and* auxiliary data from a sample or population (or "domain") or time period *other* than the one of initial primary interest. That is, auxiliary information is combined with information based on the data generated from the sample survey of the population and time that was the initial primary focus. The combination process usually depends on a statistical model that links the auxiliary information with information obtained on the population of initial primary interest.

For example, one might combine a national estimate of the incidence of student fights based on a NCES Fast Response Survey with state data on variables that may be related to fighting, such as urbanicity level, income, and so on. The combination would be based on a model of the purported relationship among the variables. The result is a *small area estimator* for the incidence of fighting at the state level. It is a special case of domain indirect estimators.

Or, one's interest may lie in updating a survey indirectly, using a *time indirect estimator*. For example, a survey run periodically, such as the Reading NAEP, might be combined with auxiliary data, based on a model of the relationship between the reading scores and auxiliary data, to estimate reading ability in a future time point between two points at which primary data are obtained and direct estimators are constructed. More specific definitions of domain indirect estimators (such as a small area estimator) and time indirect estimators are given in the introduction to OMB's *Statistical Policy Paper #21* (1993).

Precedent

NCES does not have a program to produce domain indirect estimates, time indirect estimates, or time/domain indirect estimates. Apparently, the National Center for Health Statistics (NCHS), Census, the Bureau of Labor Statistics (BLS), and the Department of Energy also have no special program as of this writing.

NCES, however, has been aggressive in building on and surpassing other statistical agencies as the need for a product appears and as resources change. The invention of licensing agreements, the use of CD-ROM for distributing data, and the use of videotapes, among other related NCES activity, illustrate the theme. Thus, it seems sensible to consider that NCES take advantage of recent developments in indirect estimation, so as to improve its surveys.

Empirical Examples Apart from Education

Breslow's paper (1989) in the *Sesquicentennial Proceedings* and the OMB (1993) report contain good illustrations of indirect estimators in various sectors, including health. Consider a small example: the NCHS national surveys of health are not directly generalizable to states. Nonetheless, there has been pressure on NCHS to produce state estimators; no resources have been provided for direct estimators. The NCHS has tried to develop reliable indirect estimators by doing the following.

NCHS obtains nationally reliable health statistics for certain subpopulations. e.g., gender, income level, and race. States per se are not in the subpopulations defined as important in the survey design although the demographic variables are.

The U.S. Census Bureau produces mid-decade estimators of the number of people within each state who belong to the subpopulations such as gender, income level, and race. The U.S. Census information is, in effect, auxiliary data that can be used to construct a NCHS indirect estimator of health characteristics for a state.

Combining the NCHS data with the auxiliary data from the Census Bureau's mid-decade data requires a model. The form of the model used to produce an indirect estimator can vary. The simplest form says that the state's health is a simple summation of the proportion of people in the state who are members of the subpopulation (i.e., female/male, high income/low income, and so on from the Census) times the mean health state of the respective subpopulation estimated from the national-level data (from NCHS). This estimator, as described, is a basic synthetic estimator. Malec (1993) describes it and others that are more complicated. The latter try to exploit auxiliary information at higher or other levels of aggregation. Citations are given for PHS reports on indirect estimators for the states, on physician visits by the disabled, based on national and regional direct estimates.

The basic synthetic estimator is being tried out by Folsom and Liu (1994) to produce state-level substance used prevalence rates based on the National Household Survey of Drug Abuse (NHSDA). NHSDA is a national probability sample survey in which it is possible to link an individual's response to characteristics of the area in which the individual lives. These include census tract/block-level information within the county on, for example, median household income. They include county-level information on arrest rates from the Uniform Crime Reports. Within a state, then, individuals' dependence on illicit drugs, for example, is regarded as a function of the block/tract level within the county and county-level auxiliary statistics and with person-level data collected in the NHSDA. The model is based on Breslow and Clayton (1993).

The result is a predicted probability of dependency that takes into account the person, his or her block/tract characteristics, and county characteristics. A probability for each arrangement of characteristics (i.e., a person) is then computed. The prevalence rate is the sum of probabilities each being weighted by the number of individuals with those characteristics living in the block/tract, county, and state. This number is itself a forecast based on the 1990 Census updated to 1992.

Evaluating Indirect Estimators

There are several ways to evaluate indirect estimators, depending on the particular form and function of the estimation. None are perfect of course; some are less ambiguous than others.

The most straightforward of these involves 1) pretending that certain data are not available when in fact they exist, 2) building an indirect estimator, and then 3) comparing the indirect estimate to a direct estimate based on the actual data. For instance, the National Center for Health Statistics has tried to construct domain indirect small area estimates of state-level mortality rates for motor vehicle accidents. The validation is against actual rates computed directly from universe data at the state level. Similarly, indirect estimates of work disability have been compared to direct universe estimates from the 1970 Census (Malec 1993).

The basic idea is that the particular indirect estimator is judged against some known value of the direct estimator. If the results agree, this fosters confidence in the indirect estimator as a possible substitute for the direct estimator.

For instance, if NCES found that an indirect estimator can be shown to produce a good estimate for 1995, based on 1994 data, data collection for a subsequent cycle (e.g., 1996) might then be skipped. Resources could then be allocated elsewhere. Similarly, if a small area indirect estimator of, for example, state-level mathematics ability works well over a 3-year period, relative to the known value, one might then skip a 4th-year cycle of direct estimation and data collection at the state level, produce the indirect estimate, and reallocate resources.

Suppose that a straightforward direct estimator or known standard is not available. How then do we evaluate an indirect estimator? One may try to judge the latter's value in predicting some known estimator, which itself is predictable from direct estimators. One may also try to construct different indirect estimators and compare results. To judge from Malec's description (1993), it is not clear how to do this right if each of the different indirect estimators could be wrong in different ways. This warrants a bit more attention.

Malec (1993) also suggests that when the indirect estimator is model-based, then examining features of the model can help to inform a judgment about the quality of the indirect estimator. This is sensible. But developing coherent theory that leads to construction of a model whose elements also are testable is difficult. Effort in this direction nonetheless seems justified. Even small theories ought to be better developed and integrated with statistical models, in this arena and elsewhere. To the extent that different indirect estimators, based on different models and theories, invite deeper thinking about evaluating the indirect estimators, is to the good.

To summarize and extend Malec's treatment (1993), evaluating indirect estimators can involve the following:

- Comparing indirect estimators to direct estimators;
- Comparing the ability of indirect estimators to predict related direct estimators;
- Comparing different indirect estimators; and
- Examining the models and theories that underlie the models for the indirect estimators.

NCES might exploit the first two approaches. Examples are given elsewhere in this section. With a few exceptions, the last two approaches have not been examined deeply. These are where NCES might make a distinctive methodological contribution.

General Implication: Time Indirect Estimators

Some NCES data collection efforts occur annually. Examples include the yearly acquisition of school district fiscal data and the NCES investment in the Current Population Survey Enrollment Supplement. Often, NCES surveys are uniformly periodic, as is the Schools and Staffing Survey, in the sense of occurring every 3 years. Or, they are nonuniformly periodic in that the time intervals between surveys may vary, the National Assessment of Educational Progress being an example. The calendars for NCES data collection provided in the NCES publication *Programs and Plans* are particularly instructive.

Recall that a time indirect estimator exploits information from one time point to describe what we know about another. Estimates of reading ability in one year, for instance, might then be inferred from estimates of reading ability at another time. A time and domain indirect estimate of reading ability may rely on reading ability data from earlier times and on other auxiliary data.

In the abstract, time indirect estimation methods seem relevant to the NCES survey effort in several respects. One may imagine, for example, that the annual data on school district fiscal matters are not available. Rather, only data for every 2 years are available. Time indirect estimates for the imagined absent years can be developed. If the time indirect estimates accord well with the actual data, then one might consider eliminating the intervening year's data collection. That is, one reduces the burden on NCES and on the respondents. The reduction provides NCES with more opportunity to do other work.

The ability to validate the time indirect estimate is possible only because NCES now collects relevant data annually. If NCES adopts time indirect estimates so as to eliminate some data collection efforts at a given time, it would still be necessary to validate the indirect estimates periodically. Even if such estimators are deemed inappropriate now, knowing about their validity seems important for the future.

The NCES Education Assessment calendar shows that the surveys on reading have occurred every 2 years. That is, we understand reading ability in the United States from 4 assessments over an 8-year period. Imagine that in the future, NCES might conserve resources by doing the reading assessment every 4 years, instead of every 2 years. Would it be possible to produce estimates for the intermediate 3 years in which data were *not* collected? One way of thinking about this is to explore time indirect estimators. That is, having reliable direct estimates from NCES for 1990, 1992, and 1994, we imagine that the reading ability data are absent for 1990 and then exploit some time indirect estimation method to produce an estimator of reading ability for 1992. The time indirect estimator for 1992 might be based entirely on 1990 data alone or on auxiliary data in combination with the 1990 data.

More generally, of course, one might explore how time indirect estimators might be exploited in the interest of estimating and justifying an estimator 4 years or more out using current data. That is, if 2-year out data can be predicted well and the evidence for the quality of prediction is good, then one can eliminate burden to respondents and NCES.

The raw implication of all this is that the periodicity of some surveys can be altered because the results of such surveys can be characterized well. This may be possible because NCES has invested resources in data collection that permit one to understand whether the results can be characterized well.

The time indirect estimation methods might be exploited in any NCES program that obtains data annually. The annual efforts up to 1995 in the elementary and secondary arena include the following:

- Public School Universe;
- Local Education Agency Universe;
- State Aggregate Non-Fiscal Report;
- State Aggregate Fiscal Report;
- School District Fiscal Data; and
- CPS School Enrollment Supplement.

Where a universe sample or census is used to build a population frame, exploiting indirect estimates to eliminate the universe sample will be dysfunctional at worst. The functions of each of the annual efforts then need to be taken into account.

Time indirect estimation might also be exploited in reducing burden and expanding opportunities in the postsecondary education arena. The relevant annual NCES efforts include eliciting data on

- Institutional Characteristics;
- Fall Enrollments;
- Completions;
- Finances; and
- Doctorates.

Similarly, the NCES calendar for the Library Statistics Program involves annual data collection efforts that might be reduced to surveys every 2 years, e.g., on public libraries. The reporting burden and the burden to NCES might be reduced by exploiting time indirect estimators. Moreover, how well the indirect estimators perform can be evaluated because NCES has obtained data on public libraries annually since 1988, and has obtained data on academic libraries every 2 years since 1988. Again, if the survey is used to build a population frame for subsequent surveys, this would have to be taken into account.

The time indirect estimators might be also exploited in surveys and assessments that are undertaken every 2 (or 3) years, in the interest of reducing burden on NCES or respondents or both. The NCES data collection efforts that have occurred routinely every 2 or 3 years, and that then present an opportunity for evaluating the indirect estimators include, at the elementary and secondary level:

- Schools (SASS, 3-year cycle);
- School Administrators (SASS, 3-year cycle); and
- Teachers (SASS, 3-year cycle), among others.

Implication: Cross-Agency Efforts

The *OMB Statistical Report* #21 represents a small but noteworthy effort to pool expertise from different federal statistical agencies. NCES was represented in the work, as was BLS and others.

An implication of this NCES effort is that the cross-agency efforts can be important and productive. Beyond this, there is another implication. Education-related data are important in some efforts to produce dependable health statistics. County-level education data, for example, have been exploited to produce more precise indirect estimators for state rates of physician visits. This use of education data reiterates the idea that, theoretically and empirically, education variables are important in characterizing health phenomena at national and state levels. NCES produces education data. NCES cooperated and assists in this production by other agencies, such as the Census Bureau. These roles are important to pursue in the interest of better design of surveys, i.e., surveys whose results can be used by other agencies.

Implications for Specific NCES Surveys

In the following, the purpose is to educe the implications of developments in indirect estimation, including small area estimation for specific NCES projects. However, the discussion is not always well informed. The NCES *Report on Programs and Plans* was a sturdy source of information. This section covers the following:

- National Household Education Survey;
- Schools and Staffing Survey;
- School District Mapping;
- Library Statistics Program; and
- Fast Response Survey.

National Household Education Survey

The National Household Education Survey (NHES) is undertaken (roughly) every 2 years on a probability sample of households with children. Special topical supplements or components have directed attention to adult's participation in adult literacy programs, school safety and discipline, and school readiness.

Time indirect estimators might be exploited to produce statistics on the years between the biennial surveys. It is not clear that there is immediate need for such estimates. Still, learning how to construct them and evaluate them can serve NCES well if 1) a biennial survey must be skipped and the indirect estimators suffice; 2) learning about the performance of these time indirect estimators is important for judging the quality of time in direct estimators in other contexts; and 3) enhancing opportunities for NCES to decrease coverage of some questions at little cost and increase coverage of others.

It is not clear whether and how small area estimators might be exploited, based on the NHES, auxiliary data, and models. The NCES *Programs and Plans* description, for instance, does not tell whether the statistics produced on the basis of the NHES are subnational. For reporting on some topics at the local or regional level, indirect estimators may be desirable. For instance, school safety and discipline matters are arguably most interesting at the city level. If resources are insufficient to provide direct estimates of, say, household encounters with school theft at the local level, then indirect estimators might be used.

Domain indirect estimates at the city level in the school safety and discipline arena might then be based on NHES, and on auxiliary data from the Common Core of Data, the NCES supplement to the Current Population Survey, or other information. It is not clear what models can be exploited to make the linkage between national (or state or regional) data and city-level data. Nor is it clear that linkage is possible.

How to validate the small area estimators produced for school safety and discipline is not clear. But there are interesting options. They include the NCES Fast Response Survey System (FRSS), the FBI's Uniform Crime Reports (UCR), the Bureau of Justice Statistics' National Victimization Surveys, among others.

Schools and Staffing Survey

The Schools and Staffing Survey (SASS) involves a periodic national probability sample of schools, interviews being directed toward school-based respondents, subsamples of teachers, and at times, subsamples of special interest. The latter included student records offices in 1993-94. The teacher survey has included a 1-year longitudinal followup of a subsample.

The NCES Common Core of Data provides the public school universe from which part of the SASS sample is drawn. The NCES 1989-90 Private School Universe File served as the population frame basis for the remainder of the sample.

The publication *SASS by State* provides statistics on public schools that are accurate at the aggregate level for each state and for the nation. The statistics are direct estimators, based on a sample design that permits their production. It seems reasonable to suppose that state-level *indirect* estimates might be produced using national-level data based on a smaller sample than is usually drawn, information available from the Common Core of Data's universe of public schools, and models to link the two. An alternative feature to exploring this strategy is that the indirect estimators can be validated because NCES does produce direct estimators against which the indirect estimators can be compared. Beyond this, the exploration may help us to learn why an indirect estimator cannot be produced or is inadequate in this instance and perhaps in others, and if it is adequate, NCES' flexibility in future SASS surveys.

Similarly, given that state-level data are available and assuming that some states have an interest in district or city-level estimates, NCES might explore the use of these estimators for this lower level of aggregation. Although evaluation of such estimates would be difficult, at least some validation is possible if in some states the district-level samples are sufficient for computing direct estimates.

The Common Core of Data (CCD) and perhaps other information supplied by the Decennial Census, for example, may be used in any time indirect or domain indirect estimator. For instance, the CCD might be exploited in attempts to build indirect estimates of supply and demand for teachers, teacher characteristics and opinions, and so on at the substate level. Insofar as municipalities, especially large cities, are concerned about the topics, the indirect indicators may be sufficient to meet their needs.

To judge from the 1995 NCES publication *Programs and Plans*, private school data generated by SASS is direct estimator based at the national and regional level, but not at the state level. The 1989-90 NCES Private Universe File may be capitalized as auxiliary variables in constructing state-level statistics for private schools or for regional statistics beyond the four censuses now used. The 1990 census data, or less likely NCES' supplements to the Current Population Survey, might be similarly capitalized. Insofar as voucher system or relative financial supports for education are state-based and that private schools have an interest in vouchers, the indirect estimators may be of value to the users of NCES data.

SASS has been planned for 5-year intervals. We then have direct estimates every 5 years on topics that are regarded as important by the users of NCES data. Suppose that estimates of statistics on the interviewing years would occasionally be valuable to states, municipalities, or the Congress or to other admiring users of NCES data.

Estimates of statistics on each year in the 5-year interval between SASS might be based on time indirect or domain indirect estimators. Suppose, for instance, that a new administration's interest lies in the best estimate possible of how many teachers will quit their jobs and form a new business, or how many teachers here engage in a new business while they maintain the teaching job. It is conceivable that indirect estimators can be constructed. Again, their defensibility is not clear.

Finally, consider the question: "Is it possible to generate indirect estimators for the state level?" Because direct estimators are available, the indirect estimates can be evaluated against them. The question may be worth addressing in a program of research on indirect estimators partly to better understand their general performance. It may be worth addressing so as to provide more flexibility in the design of SASS. That is, indirect estimators may be adequate for some parameters; and direct estimators may be essential for others. Learning which is which seems important.

School District Mapping

Regarding School District Mapping, the NCES (1995) publication *Programs and Plans* tells us that the U.S. Census identified blocks and has mapped them onto local and state education jurisdictions in 1970, 1980, and 1990. This mapping project is, in principle, important for indirect estimation at the local level. It provides a deterministic geographic link among jurisdictions—national, state, substate. The information however, may not be sufficient for small area estimation. *Programs and Plans* tells us little about what statistics at any geographic level are produced on the basis of the mapping project, despite the import in principle, of the project. It does tell us that "NCES will provide 200 tabulations of state and district totals to each of 16,000 education agencies and each state."

Library Statistics Program

The NCES Library Statistics Program is directed at public libraries, academic libraries, and school-based libraries and media centers. Public Library statistics are based on *annual* reports on a universe of nearly 9,000 libraries since 1988. Reports have been channeled through State Coordinators. The reporting is automated through the DECPLUS system.

NCES obtains Academic Library data on the universe of accredited institutions and unaccredited 4-year institutions every 2 years. The beginning year was 1988. The reporting, exploiting the software system IDEALS and state level coordinator, is done every 2 years in coordination with IPEDS. The school-based library statistics are gathered primarily through the national *probability sample-based* Schools and Staffing Survey. Data are available from 1991 and 1994; plans for the future are not yet accessible.

Consider first the annual public library statistics program. The implication of analytic developments in time indirect estimators is that the library statistics need not be obtained annually. That is, the library data might be obtained every 2 years, rather than each year. This reduces the reporting burden on respondents and the burden of processing reports for NCES staff. If time indirect estimators are adequate, the reduction in burden is not complete. A small burden would be shifted to those responsible for producing the time indirect estimators and for producing evidence that the estimators are credible relative to some standard.

Whether it is valuable to anyone to reduce the burden of reporting for public library statistics is debatable. NCES and the states have done a remarkable job to generate a low cost reporting system.

Despite this success, exploring the use of time indirect estimators may be warranted on accounts. First, the time dedicated to reporting each year for the annual public library data may be dedicated differently every other year without real damage to annual data that are based on direct and indirect methods every 2 years.

The alternative years might be dedicated to special surveys; for instance, we might like to understand the frequency of users of the public libraries, based on library card use. Now, we might like to understand periodically which books, in a list of 100, are most exploited based on check-out lists. The point is that the time dedicated to an annual report might be adjusted. The routine report may be made every 2 years. A report on special interest may be made for each intervening year. The statistics on intervening years may be generated through the year in direct estimators.

Fast Response Survey

The Fast Response Survey System (FRSS) is usually based on a national probability sample. The members of the sample are asked for information by mail or telephone.

The sample design, in some cases, is such that regional-level estimators (i.e., subnational) are possible. Further, the surveys, at times, obtain substantial demographic information on individuals or the institutions on whose behalf individuals report.

The FRSS data collected at the national level can, in theory, be combined with other NCES data obtained at state or regional level in the interest of producing local area statistics that are of interest. For example, FRSS does not produce state or substate statistics on criminal and noncriminal disorder in schools. It produces national and regional statistics. Statistics on cities are often deemed important by Congress and by the states. Statistics at the state level are often deemed important. The subtechnology of small area estimators might be brought to bear to provide those statistics. The objective seems feasible for state rather than substate levels, but both seem worth exploring.

ADJOINING RANDOMIZED EXPERIMENTS TO OBSERVATIONAL SURVEYS: SATELLITE POLICY

Introduction

There are a variety of ways to enhance the usefulness of surveys. In this essay, one such strategy is considered. The idea is to attach controlled randomized field experiments periodically to ongoing NCES surveys.

Research policy that encourages coupling the two approaches, experiments and surveys, will make both survey data and experimental data more useful for social science and public policy, and decrease the artificial separation of the sample survey and experimentation traditions. The expectation is that linkage will occasionally reduce unnecessary debates over policy-relevant data analyses. In short, a policy that invites coupling of surveys and experiments would combine the strengths of each approach while compensating for their respective analytical and administrative weaknesses.

The following section provides excerpts of work by Blumstein et al. (1986), Farrington (1988), Fienberg and Tanur (1986), Boruch and Pearson (1988), Boruch (1975), and others. It also presents some new ideas.

Definitions

Longitudinal surveys are defined here as repeated observations of the same persons or organizations or other entities in the interest of documenting growth and change. A major purpose of such studies is to understand how individuals (or organizations, and so on) change over time. Interest may, for example, lie in the growth of children's intellective achievement and how that growth accelerates rapidly during some periods (e.g., early childhood) and accelerates less rapidly in other periods. Or, the interest may lie in variations in level of delinquent activity over some period. When based on well-designed national probability samples, such surveys are the best possible approach to statistical characterization of individuals' growth, development, and engagement in various educational and social systems. Compendia of national longitudinal surveys are given in Taeuber and Rockwell (1982) and in Verdonik and Sherrod (1984).

Randomized experiments are defined as settings in which individuals (or organizations, or other units of study) are randomly assigned to one of two alternative regimens. The object of the experiment is to estimate the relative differences among regimens in a way that is unbiased, and that permits formal probabilistic statements to be made about one's confidence about the estimates. Interest in long-term differences between what are frequently referred to as "treatment" and "control" groups are often of interest and may engender the repeated observations that characterize longitudinal or panel research designs. Collections of field experiments are listed by Boruch et al. (1978), Riecken et al. (1974), and others.

The statistical models used to analyze each kind of data usually differ. Heckman and Singer's edited monograph (1985), for instance, reviews methods of analysis but not the design of such studies. But one can develop analyses that simultaneously exploit contemporary experimental design models and models designed for common panel or longitudinal data (e.g., Boruch 1975; Fienberg and Tanur 1986, 1987b).

A Proposal for Satellite Policy

The proposal for joining experimental studies to ongoing NCES surveys may be stated as follows (amended from Boruch and Pearson 1988).

Any NCES survey study should be designed so that independently designed experimental studies can be adjoined to the survey so long as 1) the experiment is compatible with the mission of the NCES longitudinal survey; 2) the risks of disruption to the NCES survey can be managed, especially in regard to the time frame, respondent's burden, and institutional cooperation; 3) designated contractors are responsible for oversight of the process; and 4) the experiment involves no appreciable cost to NCES.

This proposal is analogous to the policies on satellite use that have been used in astrophysics. The satellite, like a longitudinal survey, has a primary monitoring mission and requires considerable resources to place and maintain. Further, scientists can obtain access to part of the satellite periodically for limited, temporary investigation of important scientific questions (i.e., experimentations).

The strategy proposed here allows the research to depend on the infrastructure of the ongoing survey as a vehicle for conducting prospective experimental studies. The proposal also extends a scientific tradition of "data sharing" in the social and behavioral sciences and education research (Fienberg et al. 1985). In particular, it requires that resources be shared: population listings and sampling frames, the organizational vehicles for longitudinal surveys, and so on, not just data.

Adjoining experiments to ongoing longitudinal surveys is likely to be feasible. However, this may occur for only for a few projects, perhaps only one every year or two, because of the difficulty of coupling a special study to an already complex survey.

Justifications for a Satellite Policy

There are several kinds of justification for adding controlled tests to such a study design, as described next.

Scientific and Statistical Rationale

The mathematical conditions under which longitudinal (nonrandomized) study will fail to yield an unbiased estimate of relative program effects are well understood. Rubin (1987) provides a basic description in the context; Campbell's and Boruch's treatment (1975) is more rudimentary.

Heckman and Robb (1985) provide elaborate description for analysis of both longitudinal and cross-sectional data in an economic context.

Despite advances in the mathematical aspects of the topic, the problem of assuring that mathematical assumptions are tenable remains. Even determining whether assumptions are met can be difficult or often impossible, especially where theory is not adequate. All approaches to estimating the effects of intervention based on longitudinal nonrandomized data depend heavily on the assumption that performance of individuals in the absence of the intervention can be estimated accurately.

The assumption is patently suspect to judge from empirical comparisons of evaluations based on longitudinal against evaluations based on randomized evaluations. LaLonde (1986), Fraker and Maynard (1987), and Maynard (1987) show how estimates of program effect based on the former have been demonstrably wide of the mark in evaluation of manpower programs.

The economist's work is recent. Early research on nonrandomized clinical tests in medicine and on randomized clinical trials showed differences in results between the two. Boruch and Riecken (1975) gave relevant illustrations.

Work by Gray-Donald and Kramer (1988), for instance, reiterates the point for research in pediatrics. Observational studies have typically shown a definite association between infant formula supplementation in hospital settings and lower subsequent breast-feeding by mothers. The inference has been that supplementation then has an important potentially negative effect. Controlled randomized tests show no such difference, reducing pediatricians' concerns about supplemental feedings in hospitals.

The point of this and other illustrations is though longitudinal studies may be useful for description of growth and change, they cannot be relied on for accurate estimates of the effects of new intervention programs, at least not in the absence of strong theory.

The implications for Chapter I evaluations based solely on longitudinal study are direct and have identified been by Smith (1988). The law's demand that Chapter I effects be estimated using only longitudinal study cannot be met without heroic assumptions about children's behavior in the absence of such programs. Such assumptions may be tolerable politically. But they are often indefensible scientifically. The implications for longitudinal study of the Program on Human Development and Criminal Behavior are related if indeed the program seeks to determine how onset of delinquent behavior and resistance can be affected by intervention. They are reiterated by Farrington (1988) and Farrington, Ohlin, and Wilson (1986) among others.

A second justification for adjoining experiments to longitudinal study is that the science and technology of randomized field tests of projects has developed more or less independent of the technology of longitudinal surveys. The intellective separation is often sufficient to prevent researchers from thinking about both in designing tests of new programs or in designing longitudinal studies of important topics. There are good scientific reasons to avoid intellective parochialism here and to understand the union of approaches when the opportunity arises.

A third scientific justification stems from the observation of Fienberg et al. (1985) that although major experiments involve collecting longitudinal data, their analysis is often based on dynamic models that were not incorporated into the design of the experiment. The failure to involve these models in design of the survey, they suggest, ultimately leads to less defensible analyses of experimental results. The argument seems sensible. But little formal research on the relative gains and costs of basing designs on analytical models appears to have been undertaken.

The scientific justification for coupling experiments and longitudinal surveys is then to capitalize on the strongest merits of each. That is, one obtains both the information produced by national probability samples—often conducted over a considerable length of time—and the information produced by smaller comparative experiments in which causal inferences are more appropriately deduced. Insofar as the experiments can be adjoined systematically to surveys, their generalizability will be enhanced.

Economic Rationale: Less Costly Policy Experiments

It takes considerable effort to mount high-quality surveys. It also takes considerable effort to mount randomized tests of policy relevant programs, more effort if we recognize the difficulty of maintaining control over selection of individuals into programs and over program operations. To the extent that an experiment can capitalize on the resources and data of a survey, the experiment becomes a less costly enterprise.

Experiments undertaken by the Broward County School Board's Department of Research (1987) are a case in point. Their experimental tests of the AIM project for youth at high academic risk capitalizes heavily on a regular system of standardized testing using Iowa Achievement Tests (i.e., a periodic survey) and the infrastructure to which regular testing was based to execute the experiments. The infrastructure was especially useful in tracking the large number of children who migrated from the original 6 schools to 18 schools (Carey Sutton, Personal Communication, November 11, 1988). In a longitudinal study, for instance, we might reasonably expect the adjoined experiment to exploit one or more of the following elements of the basic survey-based study:

- Interviewer cadre, the investments in their training, supervision, and quality control;
- Questionnaire and interview design;
- Information generated in the longitudinal study about local institutional, political, and managerial constraints and stakeholders; and
- Knowledge emanating from the survey about the structure and quality of administrative records, e.g., police records, education records.

Two kinds of local statistical data generated in surveys are often crucial to a well-executed experiment: estimates of the number of individuals relevant to a particular experimental project and estimates of the temporal flow of such individuals through various systems. So, for instance, a longitudinal study that included attention to youthful co-offenders might generate good information on their number, their geographic stability and their general geographic location or locatability. Such pipeline studies could arguably help to avoid the problem of some experimental tests in police handling of domestic violence and others (Project Review Team 1988). Such information is basic to a pipeline study that would inform the design of a experiment dedicated to preventing illegal activity by co-offenders.

It would, of course, be a mistake to depend on a survey system to inform all aspects of the design of experiments. It usually cannot help much, if at all, in understanding the ethical or legal propriety of experimental tests, for instance. Nor would a survey help to understand the obstacles to implementing a new regimen in the experiment.

The implication is that field experiments can exploit surveys done in the areas in which the experiment will be emplaced, to decrease the cost of experiments. The reduction in cost stems from capitalization on human and statistical resources and savings in time.

Prophylactic Rationale

Cross-sectional and longitudinal surveys are often used to produce evidence that they often cannot support as, for example, in addressing questions in the social sciences and public policy about the impact of social programs. The Continuous Longitudinal Manpower Survey, for instance, has been and is supported primarily on grounds that it is important for understanding the changing nature of the pool of human resources available to society.

A second justification for the Continuous Longitudinal Manpower Survey is that it is useful to understand the effect of special programs in youth employment and job training. The second justification may be useful for rhetorical purposes (e.g., to gain political and fiscal support for the survey). But it is not always appropriate and is counterproductive insofar as the claim is exaggerated. That is, longitudinal surveys alone are usually not sufficient to estimate the effects of programs designed, say, to affect the earnings of individuals, some of whom happen to participate in the survey. Nor are these designs sufficient for making causal statements about the effects of programs in the health, criminal justice, and other areas. See the earlier remarks on scientific justifications and the reference to the Fraker and Maynard (1985; 1987) and LaLonde (1987) comparisons of program effects based on randomized experiments against effects based on data, notably the CLMS and the Current Population Survey (CPS).

In the case of evaluating Chapter I programs or others in elementary and secondary education, relying on a longitudinal study will merely continue a practice that is known to be risky. The estimates of program effect, if one follows the instruction of law, will be ambiguous at best and misleading at worst. To the extent that randomized experiments are a prophylactic to such results, and have been recognized as such in medicine and education since the early 1970s (Campbell and Boruch 1975), then such experiments ought to be considered seriously.

The Program in Human Development and Criminal Behavior has grappled with this issue (Farrington, Ohlin, and Wilson 1986) and continues to do so.

Calibration Rationale

An engineering justification for joining experiments to ongoing longitudinal surveys is that one may use the experiments to calibrate estimates of program effects that are derived entirely from the longitudinal survey (Boruch 1976). That is, the biases in estimates of program intervention that are based on longitudinal data can be assessed, and periodically corrected, through controlled experiments. Longitudinal studies are then likely to be more policy-relevant and less ambiguous with respect to biases in estimating program effects. Experiments are likely to benefit from their greater generalizability, lower costs, and more manageable administration.

As a practical matter, systematic calibration is a couple of decades in the future. Nonetheless, one can develop rude comparisons of results from both kinds of study. In the work on comparing estimates in supported-work manpower training programs, for instance, the biases engendered by relying on a longitudinal survey differ depending on whether one considers youth or recipients of Aid to Families with Dependent Children. For instance, the estimates for the impact on youth in 1979 was near zero for the experiments and minus $1,200 for the nonrandomized study. Estimates for AFDC women do not differ appreciably.

It is especially appealing to consider calibration in the case of Chapter I programs because the better parts of the Chapter I Reporting and Information System and infrastructure might be exploited. (See Reisner et al. [1982] for work up to 1981.) The comparison of estimates of program effect based on grade equivalents against estimates based on randomized tests may reveal that the former does well under certain conditions, e.g., for 2nd graders. The accumulation of experience about when each type of estimate is in accord can help us to understand when experiments are not needed.

Methodological Rationale: Better Methods and Data

Some of the methodological reasons for joining experiments to longitudinal studies are implicit in the earlier remarks. The economic rationale, for instance, carries the implication that experiments can be better designed so they cost less. The statistical and calibration justifications also accord with methodological interests.

The methodological rationale for joining experiments to longitudinal study can be narrowly construed, and often is, to understanding how to reduce measurement error in tests and interviews. Understanding how to elicit accurate information from people in the face of poor memory, difficulty in understanding questions, and reluctance to provide responses seems important. The problem has, at times, prompted the design of experiments in the general context of longitudinal studies.

Malvin and Moskowitz (1983), for example, undertook randomized experiments to understand how to better elicit information from junior high school students on their drug use and attitudes. The work involved comparing completely anonymous responses to ones in which identification was elicited but privacy assured by the substitute teachers responsible for administration of questionnaires. The biases reported in identified questionnaires appear to the authors to be very small except for current use of drugs.

The Weis (1987) review of research on reliability of reports on delinquent and criminal behavior suggests that new methods of eliciting information do often not work better than high quality conventional ones. The paper is persuasive on this account. Still, need to improve quality invites attention to better controlled tests. Some of the tests can be adjoined to longitudinal study.

Mathiowetz (1987), for instance, mounted studies to understand how to better ask questions about the unemployment spells of employees of a large company partly to improve quality of data in the Panel Study of Income Dynamics (Mathiowetz and Duncan 1984). Her object was to ask questions in two different ways to determine which yielded more valid results: validity standard and available company employee records. Although in this case the same sample was asked both kinds of questions, an experiment could have been designed to achieve related ends.

Policy and Political Rationale

A longitudinal study's usefulness to policy lies partly in its capacity to show change. A national shift in school truancy level may, for instance, direct attention to the problem.

Consider then that the scholarly and policy use of longitudinal data is high soon after a first wave of measurement. The use tapers off rapidly until the next wave. Consider further, several waves of measurement may be characterized by little change in the phenomenon of interest.

The implication is that "surprises" in the sense of new understanding over time will be infrequent and will decay. If they occur at all, they will be tied to frequency of measurement and frequent change. To the extent that this is true, one might choose to measure frequently. This may make possible results that show, for instance, that only 10 percent of the individuals involved in high crime commission rates in one year are involved in low or zero rate in a subsequent year. This finding has implications for policy: the high rate individuals are not durable in their enterprise and so perhaps one ought to invest in prevention rather than punishment.

It is safe to assume that such surprises will be infrequent. And the longitudinal study may have to be refreshed, in the interest of generating understanding that is not obvious.

To refresh and invigorate the longitudinal study, it seems intellectually justified to consider joining policy experiments to the enterprise. That is, one guarantees surprises—new understanding of a policy-relevant kind—by doing controlled experiments that are designed to inform policy. The regimens tested are, of course, unknown with respect to their effectiveness. On this account they also assure new understanding.

Consider, for example, Chapter I program evaluations. The expectation of some observers, to judge by P.L. 100-297, is that such programs will indeed affect truancy. A national longitudinal study may detect no effect of a program on truancy simply because a national study cannot measure as specifically, frequently, and reliably as is desirable; nor is it reasonable to expect that despite the enormous variation in such programs all will be directed toward truancy. Controlled tests of programs that replicate what appear to be the best of the *existing* programs might then be undertaken in sites that do not have such programs.

In the case of the Program in Human Development and Criminal Behavior, one might also refresh the longitudinal study periodically by undertaking experiments. For instance, handling of students at risk of further truancy varies a great deal. Ethnographic studies of the sort implied by Cooley (1988) may help to identify how most schools handle the matter and how the most conscientious do so. Designing formal programs based on what *appears* to be the best and testing these in a variety of settings is likely to be at least as important, more important perhaps, and as newsworthy as a longitudinal finding that "truancy is associated with delinquency and subsequent crime."

Related Research Policies and Origins

Precedents exist for coupling prospective methodological experiments to ongoing surveys. The Bureau of the Census, the Social Security Administration's Office of Research and Statistics, and other agencies have undertaken experiments to assess the validity of information reported to them. Measurement error and validity studies have, for example, preceded or been adjoined to the National Longitudinal Study of the Class of 1972 and the Adult Literacy Survey. In the social scientific community, the general Social Survey, which regularly employs split-half designs to study such phenomenon as the effects of question ordering.

More pertinent here is a recent effort to evaluate the USDE-sponsored Even Start programs. The program directs attention to family literacy and support services for preschoolers. Robert St. Pierre (1993) and his colleagues executed randomized experiments in five purposely selected sites to assess the relative effects of the programs. Alongside its effort, a National Evaluation Information System (NEIS) was exploited to provide information on each of a much larger number of Even Start sites to provide another estimate of program effects. Estimates, incidentally, differ on outcome measure and reasons for the differences are being explored.

Earlier precedents exist. Fraker and Maynard (1987), for example, reported on comparisons between controlled experiments and selection model-based analyses of survey data in evaluations of manpower training programs.

The proposal adjoining experiments to longitudinal surveys is related, of course, to piggybacking in observational surveys, i.e., adding questions to a questionnaire to meet the special needs of sponsors or the public. It is related also to the common practice of augmenting samples to investigate special groups that cannot be explored in a conventional national probability sample. The sample augmentation procedure of the National Assessment of Educational Progress, for example, permits states to add respondents within their states so that confident statements can be made about the state's students' achievement test scores, statements that would have not been possible with the survey's national sample design.

The satellite policy proposed here differs from earlier policies and precedents in that it suggests that the studies adjoined to the survey be *prospective randomized tests* of programs, substantive program variations, or their components. Such studies are not designed primarily to inform the methodologist; that aim is important but secondary here. Rather, they are designed to help understand what works better. The distinction is an important one insofar as social experiments engender problems that are not encountered (or are encountered in less extreme forms) in methodological experiments.

The proposal for joining experiments to ongoing longitudinal surveys has origins in the debate among scholars and bureaucrat-scholars about how much one can depend on longitudinal data. It shares an interest with those who have discussed the issue of combining experimental and sampling structures (Fienberg and Tanur 1986; 1987b). There is no doubt about the need for such data for understanding change. The debate lies in whether these data can be used sensibly to understand the causes of change.

Making comparisons between results of controlled tests is sufficiently important to evaluation policy in AIDS prevention that the National Academy of Sciences urged that agencies such as the National Science Foundation sponsor research on the topic (Coyle et al. 1991).

The National Research Council's Panel on Criminal Careers makes longitudinal study paramount in its proposed research agenda (Blumstein, Cohen, Roth, and Visher 1986). Randomized field experiments are considered generally in the context of longitudinal study as a device to test hypotheses emerging from such study and to test projects in prevention, criminal career modification, and selective incapacitation. Specific linkages between each approach to understanding are implied but not discussed in detail.

Similarly, the National Academy of Sciences' Committee on Youth Employment Programs examined major studies to understand whether one could draw firm conclusions about program effects from earlier research (Betsey et al. 1985). The committee concluded, among other things, that longitudinal surveys are no substitute for randomized experiments when the object is to estimate the effectiveness of new youth employment programs. Moreover, the committee urged the use of randomized experiments for this purpose; a satellite policy is discussed in an appendix to its report.

The proposed guideline for coupling randomized design to longitudinal surveys can also be traced to a technical advisory committee for employment program evaluation appointed by the Department of Labor. The DOL sought to learn whether analyses of manpower programs based on conventional longitudinal surveys against estimates based on randomized trials. The conclusion of this exercise was that the two estimates are not always in accord. Indeed, they differ remarkably.

The justification for the coupling of longitudinal, cross-sectional and other surveys with randomized experiments appeared in the early 1970s. In particular, the Social Science Research Council's Committee on Experimentation as a Method for Planning and Evaluating Social Interventions devoted considerable attention to the problem of generalizing from experiments.

The Committee produced two state-of-the-art monographs: Riecken et al. (1974) and Boruch and Riecken (1975), as well as a variety of papers. One of these papers concerned the coupling of randomized experiments to "approximations to experiments," such as longitudinal surveys and the models used to underpin their analyses (Boruch 1975).

Proposals for adjoining experiments to longitudinal and some cross-sectional studies have since this early work been presented formally to policy boards responsible for enhancing databases and their utility. The groups include the Policy Advisory Board of the National Center for Educational Statistics (1982), the Policy Advisory Board of the National Assessment of Educational Progress (Boruch and Sebring 1983), the National Science Foundation's Human Resources Division (1982), and others.

Examples of the Contexts to Which the Satellite Policy Is Relevant

To illustrate the kinds of setting to which the proposal is pertinent consider some examples. In what follows, different longitudinal studies and different experiments are considered. The settings bear on out-of-school youth and young adults, high school students, and children in early grades who are at risk.

Chapter I Evaluation

Consider Broward County's AIM project as a possible model. The project was targeted at 2nd graders at risk of academic failure. Risk was determined by the students' performance below the 26th percentile on the Iowa Test of Basic Skills. The AIM program involved random selection and assignment of these students to all-day programs in small classrooms, with an emphasis on basic skills; the classes were being taught by specially selected teachers.

The project was undertaken in a district that has considerable standardized testing and a Research Department that is active. The experimental field test of the AIM project exploited the testing and research infrastructure in several ways that can be emulated in evaluating Chapter I programs.

- Candidates for the program were identified on the basis of regular testing, i.e., low ITBS scores;
- Impact of the program was based on the ITBS administered to project participants and comparison students;
- Routinely collected administrative records on absences and behavior problems were used to understand implementation and outcome;
- Specialty tests were developed to capture localized differences between the randomized AIM and non-AIM students; and
- The administrative system for tracking students was used too.

Not all school districts are interested in improving programs in ways that are testable, of course. Not all schools have sufficient numbers of students at risk to justify the investment in either program innovation or formal test. Broward County School District is, for instance, the largest in the country.

The implication is that not all districts with Chapter I programs are capable, much less willing, to emulate such tests. Nonetheless, the Broward experience can help to inform the work of others, and to inform the way we think about coupling experiments to surveys and to routine administrative and academic information systems.

Multicohort-Multicity Longitudinal Studies of Delinquent Behavior

Consider surveys currently being designed by the Program on Human Development and Criminal Behavior. These surveys are relevant to proposals for Chapter I evaluation in the sense that both studies are longitudinal in character, are likely to focus on at least some common outcome variables such as truancy, and will be national in scope.

It is not hard to identify potentially interesting experiments that might effectively exploit a longitudinal study infrastructure and be worth doing. In fact, the number of options is sufficiently great to make choice difficult. The feasibility of any option may then be the determining factor, e.g., willingness of the site's public service agencies, such as police departments or courts or community-based organizations to cooperate.

For example, relatively innocuous and small but useful side experiments might be adjoined in all longitudinal studies to determine which methods are most effective locally in eliciting cooperation in the main longitudinal study or in improving the accuracy of reporting on delinquent or criminal activity. A strategy that comports with this aim might simply replicate and improve earlier experimental tests of such methods, such as the following:

- Malvin and Moskowitz (1983) on drug attitudes and use among junior high school students;
- Goodstadt and Grusen and others on the use of randomized response and other methods for eliciting sensitive information (Boruch and Cecil 1979);
- Bradburn and Sudman (1981) and others on alternative methods of interviewing and questionnaire design to improve data quality; and
- Potentially useful experimental tests are implicit in Weis (1987).

For adolescent or in-school cohorts, it may be desirable and feasible to design and test programs based on a variety of theoretical perspectives. Differential association theory (Ohlin 1988), for instance, suggests that association of target adolescents with others who are more or less delinquent will affect the targets' delinquent behavior. To the extent that school-based programs (e.g., that focus on unacceptable social behavior) or programs that attract individuals who are out of school into employment or other programs are worth testing, the longitudinal infrastructure will facilitate such testing. The extent to which shifts in association can be controlled at all seems worth testing in a controlled education, sociological, and training context.

Taking this idea further, Reiss (1986) reviewed available research on co-offenders generally. He endorses the idea put forward by Klein and Crawford that external sources of cohesiveness of gangs, if eliminated, would lead to gang dissolution or degraded cohesion. He recognizes that conventional approaches, e.g., incapacitation and social work attention, do not reduce internal cohesion and, on the contrary, may increase it. The options that are explicit in the Reiss paper and that lend themselves to experimentation include the following:

- Court-oriented efforts to sanction co-offenders in ways that are different from sanctioning individuals (to increase sense of risk), e.g., early sanctions to all co-offenders;
- Interventions designed to reduce external sources of cohesiveness (e.g., threats from gangs, revenues from drug sales); and
- Intervention designed to disrupt recruitment of co-offenders.

Consider now a different kind of coupling, one that involves a randomized test, a time-series analysis, and longitudinal study. The idea of combining these has precedent in at least one major economic effort: the Experimental Housing Allowance Program. In EHAP, poor families with certain cities were randomly assigned to various kinds and levels of housing allowance (e.g., for home repairs). In other cities involved in so-called saturation experiments, the providers of housing were given federally subsidized support to understand how to enlarge the supply of quality housing for the poor; the estimated effect in these projects was based on time-series analyses.

Related kinds of couplings have been planned but not executed in Wisconsin. Irv Garfinkle and his colleagues have begun randomized experiments on better ways to extract child support from delinquent fathers. And to understand how communitywide interventions affect such payment, saturation tests have been designed for county-level implementation. It is conceivable that similar randomized tests and nonrandomized time series or panel analyses can be executed in other areas, in the interest of understanding how to assure that young, out-of-home fathers provide financial support to their children.

Alex Weiss (1988) has considered the merits and shortcomings of randomized experiments on police handling of crime. His stress on the use of time-series approaches suggests a coupling of the approaches. So, for instance, if the general effects of delinquency deterrence are plausible at all they ought to emerge from communitywide programs that focus on norms, associations, handlers, sanctions, and so on. And in some geographic areas, pertinent saturation experiments that exploit time-series *or* longitudinal data may be feasible. Elsewhere, deterrent efforts that focus on offenders and co-offenders might be designed and tested in randomized experiments that also include long-term (longitudinal) followup.

Consider the NLS-72, HS&B, and NELS:88. These surveys are costly and widely used by the educational research and policy community. They are sponsored by NCES and have led to a variety of provocative reports, e.g., Coleman et al. (1982).

There are a variety of reasons why such studies are relevant to proposals for a Program on Human Development and Criminal Behavior. To the extent that the Program or Chapter I evaluation will involve study of the onset and resistance of delinquency among in-school children, the NCES longitudinal studies might be augmented to focus on the high risk geographic areas and people that are of primary interest. Questions might be added to ordinary questionnaires to add to the fund of knowledge.

More to the point, consider that the Program in Human Development and Criminal Behavior may be in a position to augment not its own longitudinal survey, but future NCES surveys or waves of measurement. That is, if the program invents, extends, or facilitates the invention of programs that reduce delinquency among high school students, then the Program's interest in testing them could drive the tests beyond its own borders. The drive may stem from inadequacy or irrelevance of its own target samples, or from simple interest in better use of institutional resources.

For instance, differential association theory explored by Ohlin suggests that an individual's resistance to crime results in part from a change in associations, notably a change from criminal associations to noncriminal. Inducing and maintaining such a change may involve jobs, military service, or other special handling methods. Programs designed to do the job should take account of history in locations, number of those at risk, level of risk, and so on. Information about these are available or can be collected at marginal cost from target areas in a national NCES survey. Further, the relations between the survey and local sites are sufficiently good to consider providing opportunities to do side experiments on effectiveness of such programs.

The example implies a link between delinquency research and educational research. Why would a federal office of educational research and statistics benefit from an explicit satellite policy more generally? There are several reasons. First, issues of data and resource sharing have emerged often during meetings of advisory committees for HS&B and NLS, and it seems reasonable to expect their reoccurrence. It then seems sensible to develop a program of joining experimental studies to these surveys that would help such committees and their staff understand how to respond to these issues equitably and efficiently.

Beyond this, it is not difficult to identify major survey-based studies and related multisite controlled field experiments. For instance, NELS:88 is being used to try to assess the effects of precollege programs, such as Upward Bound, on persistence in school, college applications, and so on. A series of controlled experiments on Upward Bound (Trio more generally) are being run independent of this. NLS-72 and NELS:88 have been used to study the effect of various factors on dropping out. There are over a dozen controlled experiments in the field designed to understand whether USDE dropout demonstration projects work.

Employment and Training

Let us suppose that randomized trials of employment and training programs are not always appropriate or feasible. Suppose further that there is some interest in learning from such trials, especially through using longitudinal surveys as a vehicle for their implementation. How might such experiments be carried out? Several strategies may be appropriate, and were reflected, for example, in early plans to evaluate programs of the Job Training and Partnership Act (Bloom et al. 1987). All of the following discussion assumes that experiments can be conducted in a way that permits one to take advantage of the longitudinal data and the organization structure used for its collection without disrupting that process.

Specific components of full programs may warrant testing. For example, we know very little about when, why, and how different varieties of job counseling "work." Mounting experiments in a selection of sites to assess the effects of the components of an employment and training program will often be more feasible and perhaps more appropriate than national trials on full-blown programs. See, for example, Bickman (1985) on assessing preschool programs for children in Tennessee.

Augmenting the existing employment and training regimens may be feasible in some sites. For example, how "residential" does residential training have to be? We know that some residential programs work (e.g., the Job Corps). We do not know how brief the residential experience can be while continuing to be effective (see, for example, Betsey et al. [1985] on such programs).

There is little good evidence to help answer the question "Does it 'pay' to treat the most needy, rather than the least needy?" The most "trainable" people (i.e., those most likely to benefit from training) often lie at the margin of need. And this margin often defines a population for which randomized trials are likely to be most feasible. Randomization at the margin can be coupled with other designs as well, e.g., regression-discontinuity (Riecken et al. 1974).

Selecting only the best of an array of research sites that are capable and willing to conduct experiments will not give fair estimates of the impact of programs. But such sites will demonstrate the best that can be done, thus providing evidence that may be sufficient for purposes of making policy and producing research that is heuristically rich for the social sciences.

Probable Issues and Options

The idea of adjoining field experiments periodically to longitudinal surveys is not new. But it has not emerged often and this accounts perhaps for the scarcity of thoughtful papers on the topic. Another reason for this scarcity may be the difficulties of executing the idea.

Some of the difficulties are resolvable given the current ability of research managers and manager researchers. Others require more thinking and perhaps pilot tests.

The following considers issues and options that are general, i.e., not depending on whether the experiments are adjoined to an existing longitudinal study or to a proposed study. Respondent burden is important regardless of design for example. It also treats issues that depend on whether the experiment is adjoined to an existing study, e.g., proprietary interests, or to a proposed one.

Standards for Joining Field Experiments to Ongoing Surveys

The proposal put forward earlier suggested that adjoining experiments to a longitudinal study be regarded as a legitimate research as long as

1) the experiment is compatible with the mission of the longitudinal survey;
2) the risks of disruption to the survey can be managed;
3) designated contractors are responsible for oversight of the process; and

4) the experiment engenders no appreciable cost to the agency supporting the longitudinal research.

Adhering to these standards is likely to reduce or eliminate obvious problems.

Still, one must decide which of a variety of potential experiments should and can be adjoined to the longitudinal study. Greenwood's draft paper (1988) lays out five criteria that help in making a choice. Paraphrased, the criteria include the following:

1) theoretical importance of the program(s) proposed for experimentation;
2) empirical evidence for the worth of the program(s);
3) "amount of difference" between proposed regimens and current practice;
4) compatibility with the longitudinal design; and
5) political feasibility.

The fourth item of course is part of the Boruch-Pearson (1988) proposals. Discussions and criteria for understanding political and managerial feasibility are important and have been given in, among others, Chelimsky's edited volume (1985) on evaluation at local, regional, and federal levels of government, and in Riecken et al. (1974) on managerial, ethical, and institutional and political issues, engendered by social experiments.

Greenwood's second criterion implies that evidence ought to be available from quasi-experimental or other randomized experiments. It seems sensible, given the likely cost of mounting new experiments, the need to anticipate outcomes, and the need in most field experiments to rely on earlier pilot testing of randomization procedures, measures, and negotiation strategies (Boruch and Wothke 1985).

Criterion number three is interesting in part because one can easily argue two sides. To the extent a difference between proposed regimens and existing control regimen is small, then detecting a difference in outcome will probably be difficult and perhaps not worth the effort. On the other hand, a small change is likely to be politically and managerially more feasible than a large one.

Similarly, to the extent that the difference between proposed regimen and existing control regimen is large, differences in outcome are likely to be more detectable and the product may be useful on policy and theory ground. But the managerial problems may be difficult. The handling of this matter by Riecken et al. (1978) is to encourage some testing of extreme program levels, the reasoning being that most interventions are weaker than they are predicted to be and that effects are, if the variation is effective, more detectable (pp. 33-34).

Adjoining Experiments to Existing Surveys

Proprietary interests of researchers are important, of course. The principal investigators in a longitudinal study such as a Chapter I evaluation may be disinclined to permit another research group, such as the Program on Human Development and Criminal Behavior, to augment Chapter I samples or questionnaires because this would capitalize on the Chapter I infrastructure, expertise or ideas. It would yield no obvious benefit to the Chapter I researchers. Similarly, the major sponsor for a Chapter I evaluation, the U.S. Department of Education, may see no benefit in sharing credit for an important survey by cooperating with another federal agency, e.g., the National Institute of Justice.

Some ways, quid pro quos, to meet proprietary interests then must be developed to make satellite policy possible. The National Opinion Research Center, for instance, operates HS&B and is under no obligation to cooperate with organizations responsible for surveys or experiments in another area. Moreover, developing such an obligation through contract and negotiated agreements may be difficult. There are few precedents for interorganizational cooperative research in policy and social science research. There are none for the satellite research of the kind proposed here.

Adjoining Experiments Regardless of Longitudinal Study Type

Respondent burden is and will continue to be important. For example, if an experiment on effects of Chapter I program variations asks a substantial fraction of children in early grades in a set of school districts to respond to a questionnaire and a separate study of delinquent behavior directs other questions to the same individuals, the burden on the respondents and their guardians (who must provide consent) may be increased and be notable.

Monetary payments may offset the burden. Indeed, the experience in at least some studies of adolescents suggests that payment leads to not only good cooperation of the target sample members but to requests to cooperate from those outside the sample (Howard et al. 1988).

Monetary incentives are irrelevant if there is competition for respondents in any real sense. That is, if local rule or custom dictates that the respondent can participate in only one study, then payment by a second aspiring researcher will not be relevant.

Further, monetary payments to respondents ought not be relevant if the experiment adjoined to the ongoing survey can disrupt the survey. In this case, augmenting the basic sample targeted for survey may be the only way to obtain additional information for the experiment.

Similarly, and more important, an experiment adjoined to a survey will disrupt the results of a survey in a special sense. For example, the survey researcher requires that members of the sample encounter "ordinary" conditions. The experiment will perforce introduce an extraordinary condition, albeit for a small fraction of the sample. The experimental regimen will, if effective, then affect the estimates of prevalence for incidence that are important to the longitudinal study. Again, the only resolution to this problem appears to be augmenting the sample targeted in the longitudinal study.

Augmentation of a targeted sample to reduce individual respondents' burden then may help to resolve one problem but it generates another. If a central federal, state, or local agency dictates the permissible total number of respondents, then the tactic does not help. Paying additional respondents may do so, as might other tactics.

Feasibility and Appropriateness of Experiments

Conducting controlled experiments to plan and evaluate new programs, program variation, or components is no easy matter. This is regardless of whether the experiment is coupled to a longitudinal study.

The standards for judging their appropriateness and feasibility have been laid out elsewhere, e.g., Boruch (1985). Put briefly, appropriateness hinges on answers to questions such as the following:

- Does current practice need improvement?
- Is there important uncertainty about the proposed innovation?
- Will methods other than randomized experiments yield good estimates of relative effectiveness?
- Will results of the experiment be used?

These are closely linked to standards for ethical propriety of experiments.

The standards for feasibility hinge on answers to the following questions:

- Have standards for appropriateness and propriety been met?
- Are technical and financial and human resources sufficient?
- Is the process of the new program or variation understood, described, and capable of replication?
- Is the target group and context well understood?

Methods for addressing these questions and enhancing feasibility are discussed in Bloom et al. (1987), Betsey et al. (1985), Boruch and Wothke (1985), Riecken et al. (1974), Boruch and Riecken (1975), among others.

The human resources are perhaps most important in assuring quality and feasibility of controlled experiments. For Chapter I evaluations, it seems clear from precedent that some school districts have relevant capacity, e.g., Broward County, Florida; and Austin, Texas. Some, not all, of the Chapter Technical Assistance Centers are likely to have the expertise necessary to provide counsel to school districts on the use of randomized tests for program improvements (Reisner, Turnbull, and David 1988). Indeed, directors of TACs, such as Echternacht, constitute a resource that can be capitalized nicely in this arena.

Summary

- Longitudinal surveys based on well-designed probability samples are the best possible approach available to describing growth of individuals and change at the national level.

Such surveys often do not yield defensible estimates of the effect of intervention, e.g., Chapter I programs.

- Controlled randomized experiments are the best possible approach to estimating relative effects of interventions, program variations, and so on. They are often not feasible at the national level, however.
- Coupling controlled randomized tests to longitudinal study can provide both understandings of growth or change and unbiased estimates of what works better in more local contexts.
- A formal policy for coupling experiments to longitudinal study then seems sensible. Such a policy is analogous to research policy in satellite use. The major vehicle for generating information, the satellite, is periodically reoriented and partly dedicated to special experiments and is analogous to the longitudinal study system.
- The main justification for the proposed satellite policy for Chapter I is scientific and policy relevant: better data to inform policy about how to improve programs. The secondary reasons include: economic ones, e.g., local experiments capitalize well on longitudinal infrastructure; methodological reasons, e.g., learning about how to improve data quality generally; political reasons, notably permitting answers to several questions.
- Selection of interventions for experimentation should be guided by several criteria: theoretical import of the intervention, empirical support for its promise, propriety of a test, feasibility of implementing both the interventions and the randomized experiment.
- In Chapter I, replication of exemplary projects may meet all these criteria. The experiments may for example test new ways of sustaining parental involvement, reducing dropout rates, decreasing low grades and failures, tutoring, and so on.
- Executing controlled experiments in Chapter I projects requires resources: well- trained researchers and practitioners and support for both. Failure of some projects is likely because learning how to improve and generating evidence on it is difficult. Assuming a failure rate of 20 percent for executing the experiment (regardless of program success) is reasonable.
- Statistical characterization of the target groups (who is eligible, who gets service, and so on) is essential for design of the experiments, as is careful literal and statistical description of the processes engendered by the program, e.g., time in Chapter I variation, nature of variation. Both can be generated at least crudely by longitudinal study.
- Theory will be important in the longitudinal study to estimate effects at the macro-level. The experimental programs will, if based on similar theory, help to adjust statistical vulnerability of the longitudinal work.
- A major legislative implication of this perspective is that mandates for longitudinal study must also authorize demonstrations, i.e., implementations of new programs, variations, and components.

LINKING NCES SURVEYS AND DATA FROM OTHER SOURCES

This essay concerns linking different data sets. The main vehicles for understanding in what follows are a volume edited by Hilton (1992); *Using National Databases in Educational Research*; a paper on the analysis of multiple surveys by Hedges and Nowell (1995); material generated by NCES for the NCES Advisory Council on Education Statistics; reports generated by scholarly groups such as Boe and Gilford (1992); Board on Children and Families (1995); and others. The purpose is to educe the implications of analyses undertaken on multiple data sets, in the interest of improving the design of NCES surveys.

The minutes of the NCES Advisory Council on Education Statistics (ACES) have reflected periodic interest in linking or integrating NCES-sponsored surveys. Recall, for instance, Griffith's presentation (1992) to the ACES. The agency has also sponsored scholarly work that depends implicitly on a capacity to link data in a variety of senses. Scheuren (1995), for instance, developed a variety of provocative ideas whose value hinges on linking records, record sets, or statistical data sets. The presumption here is that the general topic will continue to be of continuing interest to NCES.

Background

The Hilton (1992) book's origin lies in a project undertaken by the Educational Testing Service to understand whether different sources of statistical information, each based on national samples, could be combined to produce a "comprehensive unified database" of science indicators for the United States. Sponsored by the National Science Foundation, the project's general goal was to improve the way we capitalize on data that bear on educating scientists, mathematicians, and engineers. The book's implications, inadvertent and otherwise, are arguably important for designing NCES surveys.

Twenty-four education databases were reviewed by the project. They included the Survey of Doctoral Recipients, National Teacher Examinations, and at least four massive longitudinal studies of high school students undertaken with NCES support. Of the 24 ostensibly related databases, only 8 were deemed worthy of deeper examination. That is, they could be "linked," in some sense, given the resources available. They included the NCES NLS-72 and NELS:88, the Equality of Opportunity Surveys (1960s), cross-sectional efforts such as the SAT, and the NCES National Assessment of Educational Progress (NAEP).

As Hilton made plain in the book's preface, the project was "not feasible." Put more bluntly, the ETS effort to combine data sets was a flop despite competent and thoughtful efforts. The databases that were chosen for examination could not be used for the purpose considered, i.e., to produce a comprehensive science database. It was nonetheless a project noble in aspiration and diligent in execution.

The questions that were posed about the available databases and which are relevant to linking any data sets, seem important for designing new NCES surveys. Put in modified term, the questions are as follows:

1) What *variables* are common to various databases?

2) What *ways of measuring* the variables, *ways of sampling*, and administration are common, making comparison (or linkage) among data sets easy?

3) What *differences* in ways of measuring, administration, and sampling make comparison (or linkages) dubious or difficult?

4) What can be done to *fix* different data sets so they are "comparable" (or linkable) in some way and therefore make it sensible to put them together?

The Hilton book contained no detailed catalog of why the databases failed to meet one or more of the criteria implied by the questions.

Hedges and Nowell (1995) attacked a different but related topic, understanding sex differences in tests of mental activities of various kinds based on disparate surveys. They chose to depend only on studies based on samples of roughly the same target populations and that purportedly measured the same abilities, e.g., reading. They selected only studies that approached questions 1) and 2) above in similar ways. Their group of studies included NCES-sponsored work, notably NELS:88, NLS-72, HS&B, and NAEP (trend data only), and Project Talent and the National Longitudinal Youth Survey sponsored by the Department of Labor.

There was sufficient commonality in what was measured on whom in the Hedges-Nowell ambit to produce an informative analysis. It is a fine illustration of combining data sets in the interest of how males and females differ on mental abilities. Moreover, the dependence on well-defined national probability samples avoided the inferential problems in earlier studies, notably depending on self-selected samples (as in SAT/ACT testing), idiosyncratic samples (for example, in test norming), and distributional assumptions (to get at characteristics of extreme scores).

Questions That Have Been Addressed

What plausibly accounts for a decline over a decade in Scholastic Aptitude Test (SAT) scores? In the Hilton's (1992) book, Beaton, Hilton, and Scharder take the 1960-72 decline seriously, based on combining SAT cross-sectional data. Their analyses suggest that a decline on account of real reduction in student ability alone is unlikely. Over the period, the number and heterogeneity of youth who took the test (from 2 to 3 million) increased. There were increases in the number of youth at the bottom of the test score distribution (from 2,000 to 54,000).

What might account for cross-sectional declines in the mean visual-spatial test scores achieved by high school seniors in 1960 and seniors in 1980? Hilton argues that a portion of the decline is not attributable to any real change in ability. Rather, he maintains that it is attributable at least partly to increases in high school completion rates during the period (from 67 percent to 74 percent) and related demographic changes. The available data evidently were insufficient to illuminate competing explanations such as changes in curriculum. Oddly, Hilton ends all this with a *non sequitur*. He said that differences in sampling method and administration are such that "what the net effect of all these may have been impossible to say. The conservative position is that they balanced each other."

Are the tests given to large numbers of students measuring roughly the same thing over long stretches of time? Based on factor analyses of test scores of 1972 and 1980 senior high school cohorts, and of scores from longitudinal testing, Rock (1992) maintains that there has been no real change in factor structure despite (unspecified) changes in ways that tests were administered and characteristics of students.

Is it possible to say much about the persistence of a youth cohort's interest in science over a 2-year period and about whether cohorts born a decade apart are similar in their persistence? Valerie Lee's (1992) analyses were based on NLS-72 and a followup of them, and on HS&B, a longitudinal study that includes a cohort of 1982 seniors. There were radical changes across the cohorts: both above-average and below-average students, in more recent years, leaned toward science and mathematics. Within the cohorts, the rate of declaring science, math, and engineering as a course of study dropped remarkably regardless of racial/ethnic category.

The analyses in the Hilton book dedicate much attention to the methodological problems of exploiting two or more databases in combination rather than to substantive research results. Consider the following:

Even in studies designed as longitudinal efforts, the structure of a question's bearing on a particular topic may change dramatically over time. This means that the longitudinal changes in the trait that is targeted by the question will be difficult or impossible to discern. Lee's paper in the Hilton book was instructive on this account.

Lee suggested that one could in principle construct a question addressed to high school students about their planned major course of study and a parallel question addressed to the same students when they reach college level about their actual major course of study. To judge from Lee's work, the investigation of persistence of students' interest in science is thwarted by remarkable differences in the way the relevant question has been handled. Multiple-response categories in one round of a longitudinal survey have been followed by open-ended questions with not clearly related coding categories in the next.

Similar changes in question format, in repeated rounds of a longitudinal survey or across different surveys, affect measures of achievement across time (unless special provisions are made for equating tests that are not comparable), attitudes toward science, and so on.

Less obviously, skip and detour patterns in otherwise similar questionnaires may differ. The result can be (and for certain studies has been) the elimination of information from one target sample/database and the production of information in another. For instance, students who said that they were oriented toward vocational education in a high school level survey were then asked to skip a block of questions bearing on college. Some of these students changed in their interest and went on to college. The loss of the block of questionnaire items on those who changed their orientation is important in its own right. Further, information available on them from a later survey round differs from that on college students who did express an early interest in college.

Whether to survey individuals who dropped out of school has differed across longitudinal surveys. Following dropouts is more common now. But the noncomparability means that some data must then be ignored, i.e., on dropouts. This means that some analyses cannot be done, for instance, on what happened to dropouts from high school in the 1960s versus the dropouts of the 1970s or 1980s.

The Hedges and Nowell study (1995) was less ambitious in some sense than the Hilton project, but no less instructive. Their focus on national probability samples helped greatly to "simplify" the task of summarizing the results of multiple studies in order to learn where men differ from women in mental abilities. Further, focusing on certain abilities that were measured in each study, regardless of how they were measured, advanced our understanding. Exhibit 1 illustrates the simplification.

There was sufficient commonality in what was measured to produce comparisons. Reading ability was assessed in all six studies in the Hedges-Nowell compass, for example. This permitted the authors to recognize that differences in mean performance level between men and women are reliable but small (women surpass men) and variance across gender differs at a low level (men are more variable than women). Mathematics ability was measured in four of the six studies. Results suggest a reliable but small mean difference favoring males and again, more variance among males than females. Despite "small" differences in mean ability and variance, of course, large percentage differences can appear between the sexes. That is, remarkably more males relative to females appear in the upper tails of distributions. Further, NAEP trend data suggest that the ratio of male-to-female variance has not changed appreciably over time.

Such results run counter to small-scale studies reporting declining difference between the sexes in ability level. Independent research show high male-to-female ratios among selected "very talented" samples. The Hedges-Nowell work suggests that the ratios are plausibly attributable to small mean and variance differences, apart from "differential selection by sex" (p. 45).

The Pedigree of Efforts to Put Different Databases Together

The idea underlying any linkage study undertaken by NCES or by others is that putting together data from different sources can help us to learn something new. The combination can help to learn something that cannot be learned from individual sources.

The idea has a fine pedigree. Alexander Graham Bell, for instance, exploited the notion in his study of genetic transmission of deafness. He depended, in the late 1880s, on completed Census Bureau interview forms found strewn in a government building basement and on genealogical records from other sources (Bruce 1973).

The pedigree of linkage studies is also reflected in contemporary efforts to evaluate social programs. In studies of manpower training, it has become common to link the employment records on specified individuals to their program records, and to link these data to research records on the individuals (Rosen 1974). In agriculture, health, and taxation, there have been fine studies of why and how one ought to couple data from different sources in a variety of ways (Kilss and Alvey 1985). From papers by Scheuren (1985) and others, we may learn about contemporary history of record linkage algorithms (e.g., developed by Tepping and Felligi-Sunter), the construction of matching rules and the information exploited in matches, the idea of linkage documentation, and various approaches to adjusting for mismatches. We can learn about the role of privacy issues and statistical analysis implications from a related body of work, e.g., Cox and Boruch (1988). We learn about appraising the benefits and costs of linkage of administrative records, or the difficulty of doing so on account of sloppy practice, from aggressive investigatory agencies such as the U.S. General Accounting Office (1986a and 1986b).

Scheuren's paper (1995) for the NCES Conference on the Future of Data Collection has a different but related pedigree line. It focused on better exploitation of administrative records in NCES survey contexts, and conscientiously exploited such records more generally. One can trace the theme to John Graunt's efforts in the 17th century to learn how to use records in the Crown's interest. Graunt exhorted the Crown to learn about the kingdom through a lens consisting of compilations of records in statistical form, on the counts of soldiers-at-arms, for instance, and the numbers of births, deaths, and so on. Scheuren, similarly thoughtful and exhortative, generates ideas and reiterates others' ideas about how to augment the administrative records and understand them better through surveys.

The title of Hilton's book, *Using National Databases*, may suggest to some readers that they can learn something about whether, why, and how massive studies are combined and used. This belief will be born of recognizing recent work on how to enhance the usefulness of statistical data. Such work has been economically oriented, e.g., Spencer's work (1980) on benefit-cost analysis of data used to allocate resources and the follow-up papers by Moses, Spencer, and others. It has been based on scholarly interest in why and how social research data, including educational and health research data, are used; Kruskal's volume (1982) is a gem on this account. The work has been deepened by serious attention to how statistical data and results are misused.

The analyses contained in the Hilton book are not burdened by this knowledge. They failed to put the ETS linkage studies into the larger context of such studies or the still larger context of design and exploitation of databases and survey. We learn about attempts to link the Armed Forces Aptitude Battery to tests given in the longitudinal HS&B survey and to SATs. But we are not told about how this would enhance science indicators or inform decisions or, most important, improve the design of surveys.

Similarly, the Hedges and Nowell paper does not consider the implication of the work for the design of better surveys that can be linked in any sense. This is despite the fact that the authors are sensitive to the implications of their work on other accounts.

Building on Efforts to Put Data Sets Together

Despite the Hilton project's considerable investment in figuring out how to put different databases together, and despite the conclusion, that the databases at hand could not be put together sensibly in the interest of science-related knowledge, the book offered little counsel on how matters might be improved. Hedges and Nowell (1995) offered no counsel either, despite what can be regarded as a successful attempt to put different data sets together to advance our understanding. Scheuren's work (1995) bears naturally on linkage, but the word and synonyms for it do not appear in this paper as it does in other products of thinking. Despite the fact that the Board on Children and Families (1995) focused on "integrating federal statistics," there is no substantial examination of what integration means and its relationship with coupling, merging, pooling, and so forth. This presents something of a challenge.

Vernacular and Definitions

The Hilton book's vernacular is sufficiently different from technical parlance in related areas to confuse some readers. For instance, there are repeated references to "linking" and "merging" of different databases. But these terms are undefined. The reader should be aware that the terms have not been defined here either. Further, the book's use of them is, at times, *not* the same as is customary in contemporary statistical work of the sort, e.g., linkage being defined as combining micro-records based on a single common identifier. At times, the book's use of the word "link" is to imply an intention to "put together." At other times, the word "link" means to stratify the units in each database in the same way (e.g., high ability, Hispanic, and so on) in order to look at how frequencies in these strata change over time on a dimension such as persistence in studying science. The word "merge" is used to describe putting different records together that may or may not have a common source.

The phrase "pooling data" was used by Hilton and has been used by others, in the sense of doing a side-by-side comparison of statistical results from each of several different data sets. This use of the phrase is not as some readers would expect. Pooling data for some analysts means combining the data from two or more samples of the same population into one that can be analyzed as a complete sample. For others, pooling means combining the results from samples of different populations.

One of the implications of this vernacular problem for NCES is that discussion, analysis, and agreement on terminology are in order. Because there has been little standardization in educational research, NCES has, in recent years, played a leadership role in getting state education agencies to agree to common standards and definitions in statistical reporting. NCES can play a related role here, and to refresh the roles taken by the IRS Statistics of Income division, the Census Bureau's methods division, and others. That is, NCES can help to make plain what we mean by

- "Combining" data sets or surveys;
- "Linking" data sets or surveys;
- "Merging" data sets or surveys;
- "Pooling" data sets or surveys; and
- "Integrating" surveys.

Absent explicit definitions, reaching mutual understandings in the statistical community will be difficult or impossible. And most important, designing surveys so they can be linked, compared, merged, and so on will be impossible. NCES can be a leading agency in this effort.

Questions

The Hilton book provides ample evidence that questions about economic status or race/ethnicity or other important topics are asked differently across surveys and data sets. Differences prevent straightforward comparison. There are, however, no recommendations about whether and how to standardize such questions. There is no discussion of how directing two or more varieties of the "same" question to respondents in a survey can help to equate or calibrate the different questions across surveys. There is no serious exploration of whether and how imputation methods can help in doing so. Yet, we know that embedding different forms of the same question in a questionnaire, for a subsample at least, is a decent vehicle for learning about relations among questions. More general tactics might be invented, based perhaps on the test-equating strategies that have been explored by Holland and Rubin (1992), among others. Certainly the matter is pertinent to NCES' investments in learning how to integrate (and in what senses to integrate) the longitudinal and cross-sectional surveys that it sponsors (Griffith 1992).

An implication of this is that survey questions need to be designed with linkage in mind. NCES often does this implicitly, and in an ad hoc fashion. We are unaware of an explicitly written standard for doing so as part of NCES survey design strategy. Nor does there appear to be a systematic program of empirical side studies or pilot work by NCES that regularly takes linkage seriously.

Analyses

In the Hilton book, there are few substantial references to multiple independent analyses of the same data sets. Hedges and Nowell are more conscientious on this account. For example, there are no references in the Hilton work to other analyses that are suspect or arguably wrong. This can be regarded as a shortcoming. It is also symptomatic of the lack of good registry for tracking who analyzed what data set.

No federal agency or private foundation, including NCES, has an excellent system for tracking the research uses to which its data sets are put. This makes the evaluation and improvement of any given survey difficult. It makes development of better statistical design very difficult.

An implication is that constructing registries of analyses is one option that NCES might consider in the interest of improving NCES surveys. More conscientious efforts by authors and journal editors to assure the proper citation of data sets is another. A third option, related to the first two, involves better exploitation of contemporary Internet capabilities to build an informative registry of analyses of NCES data sets in the interest of improving survey design. It is described later under the topic of new technology.

It is important to maintain a sense of history in this. Three of the Hilton (1992) chapters were excerpted from reports produced in 1975, 1977, and 1983. The chapters contained no discernible updating. One concerns the declines in mean reading test score based on data generated in 1960 and 1972. There was no attempt to relate the data or the analyses to more recent arguments about test score declines. This lacuna is astounding given that President Bush and President Clinton stressed an education agenda based on what were claimed to be declines in student performance, declines found to be misleading by these analysts.

Documentation

The Hilton book recognizes the investment that statistical analysts must make in learning the "ponderous user's manuals" for complex data files. But the book presents no deep thinking or data on the matter. Hedges and Nowell are also silent on the matter. This is a general and nontrivial issue. Learning how to learn easily about complex files and how to teach well about complex data files seems important.

Some attention is being dedicated to the topic, if we interpret properly the current efforts of NCES. The NCES has generated and issued Read Only Memory diskettes (CD-ROM) that introduce complex data less formidably than the way public use tapes have been introduced. Beyond this, it is not clear whether and how NCES invests resources in making data file documentation less difficult to deal with.

It seems sensible to expect those who have made distinctive contributions to the quality of documentation (for instance, ICPSR) to collaborate with statisticians in this task. At least one major contractor to NCES, the American Institutes for Research, actually does research on the topic of "readability" of documents. Work of this sort might be exploited by NCES to enhance the ease of use of documentation on its data files.

Naming Surveys

It may not seem difficult for some readers to keep in mind the eight studies that are used in the Hilton book. But it is for this writer. The difficulty lies partly in the disconnectedness of the book's chapters. The difficulty goes well beyond the book, and is partly numerical. The multiple pieces of any given survey must be kept in mind. One or more of five points in time in the NLS-72 may be a focus of study. Any one or more of three points might be exploited in the NCES HS&B surveys.

Part of the difficulty may also lie in our predilection to name rather than to number. It is more pleasing, perhaps, to talk about "High School and Beyond" or HS&B than about survey #8.2, just as it is for our Chinese colleagues to refer informally to the "Red Flower" factory instead of Factory #26.

The implications of this "naming" problem for NCES are not clear. There is sufficient opportunity for confusion or difficulty to argue that a name such as "NLS-72" is more informative to many potential users of data than "High School and Beyond." It seems reasonable to argue that "Wave 2" is an important amendment to study, e.g., NLS-72: Wave 2. Perhaps this is as far as we can go.

Missing Data

Missing data are ignored by analysts in Hilton's book, chapter one, by Valerie Lee. Nor was the topic mentioned in works that are at least as important, by Hedges and Nowell (1995), Board on Children and Families (1995), and Boe and Gilford (1995).

This is despite the fact that reasons for missing data and the models that might be used to impute the missing data can differ across databases, just as definitions, sampling methods, survey conditions, and so forth differ across databases. More to the point of some analyses, missing data or differences in the reasons for it are not considered in understanding whether data from different sources can be sensibly compared. See, for instance, Little and Rubin (1987) and Rubin (1987) on imputation. At bottom, this suggests that another criterion be used by NCES to make judgments about the possibility of linkages among databases: missing data.

Major Factors

Various chapters of the Hilton book remind the reader to take into account both obvious and subtle factors in using the results from different surveys that might be thought comparable: differences in the definition of the target population, sampling frame, selection of organizational units, selection of individuals within units, cooperation rates, conditions of administration, coding of open-ended responses, multiple response categories, and timing of measures. Three major multimillion dollar surveys, arguably more, have differed notably in all respects, making comparison very difficult. Yet the book offers no advice on how to better structure the portfolio of longitudinal or cross-sectional surveys sponsored by the federal government.

Understanding how to design a portfolio of longitudinal and cross-sectional studies so that they *can* be put together (compared, linked, coupled, yoked, or otherwise used) goes well beyond what the book's authors tried to accomplish. With the exception of a chapter by William Turnbull, a statesman in the educational measurement arena, and one by the editor, Hilton, they confined themselves to the tasks at hand. Since the time that they engaged in the enterprise, NCES appears to have tried to make progress along related lines.

NCES sponsors an astonishing variety of longitudinal and cross-sectional surveys, at least four of which are exploited by the ETS Project. The agency began to collect longitudinal data in 1972, initiated six longitudinal studies afterward, has been asked by the Congress for more, and has supported a large number of cross-sectional surveys. The problems of how to develop an integrated portfolio of studies, and what integration means, how to integrate, in the face of disparate demands from Congress and the educational research community, and others under the influence of severe limitations on staff size, and other factors, are formidable.

In a sense, the Hilton book helps to understand and to justify what NCES has done to integrate studies (in the NCES vernacular) if not to create "unified databases" (the ETS parlance). One NCES initiative, for instance, focused on identifying rationales and settling on a rationale for integration, and for shaping the relations among longitudinal surveys and the relation between these and cross-sectional surveys (Griffith 1992). This does not differ in spirit from the book's focus on longitudinal study as a vehicle for a unified database but gets well beyond it. The NCES focus, to judge from Griffith (1992), is on universe and sampling frames and on how to develop agreement on each, in the interest of integration, for the design of new surveys. Hilton and his colleagues make plain that their difficulty in developing a unified database on science indications was attributable to differences in each factor.

The Hilton book alludes to factors beyond sampling that may influence the construction of unified or integrated databases. But there is no pursuit. For contemporary work at NCES and perhaps other statistical agencies, the questions are numerous and the search for answers serious. Which particular surveys are sensible targets for integration out of the portfolio of all surveys that have or might be done? How do we decide? For education surveys undertaken by NCES and others, what should be the grade span of surveys, the time between rounds, the survey's lifetime, the time between initiating new cohorts, the starting grades of cohorts (Boe and Gilford 1992) What rationale based on integration can inform the choices? How can an integration standard influence surveys on the allegedly crucial transition periods from kindergarten to preschool, middle school, and so on and durable policy issues such as the supply of science-oriented students and teachers?

These questions deserve wider attention from the statistical methods and policy communities and the disciplinary communities with which they collaborate. Here again, there appears to be fine opportunity for thinking at the National Center for Education Statistics and other federal agencies (not just statistical ones) and groups that advise them, such as the National Academy of Sciences and the Social Science Research Council.

EXHIBIT FROM HEDGES AND NOWELL (1995)

Table 1—Summary of the characteristics of the six data sets

Characteristics	NLS-72	NLSY	HS&B	NELS:88		NAEP
Year of assessment	1960	1972	1980	1980	1992	1971-92
Sample size	73,425	16,860	11,914	25,069	24,599	Varies
Population	All 15-year-olds	12th grade students	Non-institutionalized 15- to 22-year-olds	12th grade students	8th grade students as of 1988	17-year-olds in school
Abilities measured						
Reading comprehension	◆	◆	◆	◆	◆	◆
Vocabulary	◆	◆	◆	◆		
Mathematics	◆	◆	◆	◆	◆	◆
Perceptual	◆	◆	◆	◆		
Science	◆		◆		◆	◆
Social studies	◆				◆	
Nonverbal reasoning	◆	◆				
Associative memory	◆	◆		◆		
Spatial ability	◆			◆		
Mechanical reasoning	◆		◆			
Electronics information	◆		◆			
Auto and shop information			◆			
Writing						◆

NEW TECHNOLOGY

Introduction

The object here is to describe how NCES might use the Internet and the World Wide Web (Web), the Internet's graphical component, to improve the design of surveys. The main vehicle of illustration is George Terhanian's Home Page (HTTP://www.dolphin.upenn.edu/~terhania/). It relies on subtechnologies available to NCES that can in turn be exploited to improve NCES survey design.

Definitions: What Does It All Mean?

The Internet and the Web have spawned a large, somewhat confusing, vocabulary. It is necessary, therefore, to first provide definitions of several terms before describing how NCES might better exploit the Internet and the Web. Providing definitions that are precise is a challenge, however, because new terms continue to emerge, and the meanings of old terms continue to evolve, as the Internet and Web expand. Consider, for example, how the meaning of "server" has changed. A few years ago, "information and file provider" would have sufficed; e.g., a server provides information and files to clients. Today, this definition seems too narrow—it does not account for the capacity of a server to receive, process, and store information (e.g., responses to questionnaire items) that clients might send.

The lack of an official Internet dictionary, no matter how inchoate some terms may seem, also makes providing definitions difficult. "Electronic mail," "bulletin board," "discussion group," "listserv," and "newsgroup," for example, all refer to slightly different methods of sharing information. But discovering how these methods differ requires perseverance: a call to a computer-literate friend, a trip to the library or bookstore, an on-line database search, and so forth. These qualifications aside, the definitions (e.g., see Howe 1995; Raisch 1994) are as follows:

Electronic Mail: A system of sending information and files to anyone who has access to the Internet through an e-mail account. Messages are automatically passed from one computer user to another, often through computer networks and/or via modems over telephone lines.

Bulletin Board: A message database where any user may submit or read any message in public areas. It is also possible to post (i.e., to place for public perusal) other types of files (e.g., statistical software) on bulletin boards.

Discussion Group: A mail system through which members exchange messages. Membership in particular groups is often based on a common interest (e.g., hierarchical models) or affiliation. Separate messages are sent individually to each member.

Listserv: A mailing list server on **Bitnet**, an academic and research computer network. Listserv is Bitnet's version of a discussion group.

Newsgroup: A combination bulletin board/discussion group. Messages are placed in a central location, for example, like a bulletin board. However, like a discussion group, access to these messages is generally restricted to the particular newsgroup's members.

Protocol: A standard, or set of formal rules, that defines the method of communication (i.e., how to transmit data across a network) among computers. There are a variety of protocols. The more popular ones include Gopher, FTP, Telnet, and HTTP.

Gopher: A user-friendly protocol that relies on hierarchically linked menus. One limitation of Gopher systems is that the client may have to work through several layers of menus before locating a desired file.

File transfer protocol (FTP): A protocol that allows for the transfer of files from server to client. Although menus may exist, those that do generally lack the detail of Gopher menus.

Anonymous FTP: A variation of FTP. An interactive service provided by many Internet servers allowing any user (i.e., those who do not possess accounts) to transfer files.

Telnet: A protocol that may permit a remote client to log on to another server. This method does not permit the client to retrieve actual files, however.

Network: Computers that use the same protocol to exchange information.

Internet: The network of networks.

World Wide Web (WWW or **Web)**: Computers that communicate via the **Hypertext Transfer.**

Protocol (HTTP): HTTP differs from other protocols in two important respects: 1) it enables clients to view graphics, and 2) it relies on hypertext links. **Hypertext** or "text that is not constrained to be linear" (Magid, Matthews, and Jones 1995, p. 8) indicates a reference to another document or file type located elsewhere. To retrieve the referenced document, one need only "click" on boldfaced and/or underlined hypertext.

Uniform Resource Locator (URL): A unique address that specifies the target (i.e., a referenced document or file type) of a hypertext link.

Hypertext Markup Language (HTML): The language of HTTP and the Web. HTML requires authors to insert a variety of formatting information or "tags" on a page of text to indicate, for example, italics, underlining, new paragraphs, links to other documents, and electronic mail addresses.

Graphical User Interface (GUI): The use of pictures rather than just words to represent the input and output of a computer program. Popular Web browsers (e.g., Netscape and Mosaic) and popular computer operating systems (e.g., Microsoft Windows) make use of GUIs.

Multipurpose Internet Mail Extension (MIME): A systematic method of categorizing transportable (via the Internet) file types. A file extension (e.g., .au for sound, .xls for Excel spreadsheet file, and so on) indicates the specific file type. Transportable types of files include images, sounds, motion pictures, word processing documents, and so forth.

How Does NCES Now Use the Internet?

Aside from sending and receiving electronic mail, NCES now uses the Internet primarily to disseminate general information, reports, and raw data. It is possible, for example, to retrieve any number of NCES-produced items through the ED Gopher server. NCES asks that users not access its servers through the "somewhat cryptic" (Davis and Sonnenberg 1995, p. 136) File Transfer Protocol (FTP) method, and denies access to those who use the Telnet protocol to access its site. Until recently, NCES used the World Wide Web only to display and describe several publications (e.g., *The Condition of Education, The Digest of Education Statistics*, and so on) available through the ED Gopher. Since mid-November, however, hypertext versions of some of these publications have also been made available on the Web.

How Might NCES Use the Internet to Improve Survey Design?

NCES might want to consider exploiting the flexibility of the Internet, particularly the Web, to create and strengthen ties in a variety of ways with those who analyze NCES data. The rationale is that a deeper understanding of the experiences of those who analyze survey data might help NCES to design better surveys. In addition, NCES might also use the Internet and the Web to elicit, exchange, and access information from numerous sources in order to educe the implications, or at least track the development, of new analytic methods for the design of surveys.

Why Focus on the World Wide Web?

The Web possesses at least five attributes that make it an attractive vehicle for eliciting, exchanging, accessing, and distributing information. First, it enables different types of computers (e.g., IBM, Macintosh) to communicate through a common protocol (HTTP). Second, Web graphical browsers (e.g., Netscape, Mosaic, and so on) are available at no or low cost for most popular operating systems. Third, these browsers are relatively easy to use (because of their graphical interface), flexible, and powerful. They can interpret documents written in HTML, for instance, as well as several types of graphics files. Moreover, in many senses, browsers transcend protocols through their ability to access HTTP, Telnet, Gopher, and FTP servers. Fourth, the latest release of HTML allows authors to create fill-out forms (e.g., questionnaires). Fill-out forms, in particular, exploit the capacity of Web servers to receive, process, and store responses. Finally, the Web is growing rapidly—by more than 500 percent in the past year (WebCrawler 1995). There are now more than 40,000 Web servers and about 10 million daily Web users (Netscape Communications Corporation 1995). The estimates are crude, however, because the Web, for the most part, is unregulated. No official registry of servers exists and many server administrators choose not to track usage (e.g., the number of visits to a home page or the number of downloads of a particular document), although they could do so easily.

What Might NCES Do? Strategies to Elicit, Exchange, Access, and Distribute Information

This section describes several strategies, some of which are related, that NCES might implement to elicit, exchange, and access information from numerous sources. It also describes strategies to disseminate information. For an illustration, readers are again encouraged to visit Terhanian's home page at: HTTP://www.dolphin.upenn.edu/~terhania/.

Strategy 1: Elicit Information Through Fill-Out Forms and Electronic Mail

NCES might want to consider eliciting information through graphical fill-out forms and e-mail from those who are actually analyzing NCES data (e.g., licensed users). The implication is that data users are an underexploited, though valuable, resource. Questions that NCES might ask include the following:

- What methods do you employ when analyzing survey data?
- What problems pertaining to the design of NCES surveys have arisen?
- Have any journals published your work?

Analysts are not the only ones from whom NCES might elicit information. NCES is obliged, at times, to ask questions of the general public that bear on data use. The commissioner of education statistics, for example, is "responsible for providing continuing reviews including validation studies and solicitation of public comment on NAEP's conduct and usefulness" (White 1994). NCES might therefore provide a Web window (e.g., fill-out form) through which the public might either ask or answer questions about NAEP and other surveys.

Although the ability to post questions on Web pages and the capacity of Web servers to collect, process, and store responses may have direct implications for the administration of future NAEP surveys, we have focused here, and throughout, on strategies that might influence the content of surveys no matter how they are administered.

Strategy 2: Distribute Spreadsheet Files Through the Web

NCES generally distributes raw data and finished products via the Internet; that is, seeds and mature trees. There is an opportunity for NCES to distribute saplings as well. This strategy recognizes and relies on the ability of spreadsheet software, notably the most recent versions of Lotus, Quattro Pro, and Excel, to hold alphanumeric data, and graphical displays based on this data, in different sections or pages of one file. By using a mouse to click on reference tabs (i.e., links) within a spreadsheet file, the user can move from page to page.

This strategy also recognizes and relies on the ability of Web browsers to configure helper applications to interpret spreadsheet files. For instance, after the user clicks on a Microsoft Excel spreadsheet file (.xls extension) located on a Web, Gopher, or FTP server, the Web browser (e.g., Netscape), because it does not recognize the .xls file extension, will ask the user how he or she wishes to handle the file. The user may instruct the browser either to save the file or to open a local viewer, i.e., the particular application (e.g., Microsoft Excel). The user may also instruct the browser to thereafter open the particular helper application automatically whenever a file with an .xls extension is selected.

Spreadsheet files are a natural home for information that NCES might receive from data analysts. Depending on the questions that NCES decides to ask, the file might reveal what analytic methods researchers have applied to the NCES data, names of journals that have published articles, titles of published articles, years of publication, and the like. NCES, through this method, can then count publications, create displays, for instance, of NELS:88 publications by year, and sort the information however it chooses. Periodically, NCES might also post the updated file on a Web page to provide others with access. This information might help current and potential researchers to shape their analyses and it might also lead to the exchange of information. "Why isn't my article there?" or "Here's another," researchers might say to themselves. And they might then send NCES a reference for their own particular article or for others of which they are aware. Or they might send an updated spreadsheet file to NCES, thereby eliminating much of NCES's data entry work.

NCES will have to choose which type or types of spreadsheet files to distribute. Although we recommend any of the most recent versions of Windows software because of their widespread use and hypertext-like tab features, it is not possible at this time to open, say, an Excel file with Lotus software because the tabs pose conversion problems. Nor is it possible to use Macintosh software to open a Windows spreadsheet file. NCES might therefore consider distributing a more generic type of spreadsheet file (e.g., an Excel 4.0 file) as well.

Strategy 3: Track the Emergence and Development of New Analytic Methods Through the Web

It is not always clear how advances in statistical theory or technology might affect the design of future NAEP surveys. But the question is important enough to warrant attention. NCES might then also use the Web to track such advances. NCES might post references on a Web page, or links when appropriate, to journal articles and books that describe or use new analytic methods, including multilevel modeling, meta-analysis and cross-design synthesis. NCES might also provide a Web window through which Web users report additional references and the Web addresses of informative home pages. The home page (HTTP://www.ioe.ac.uk/hgoldstn/home.html) of the *Multilevel Models Project* (MMP) that is based in the United Kingdom, for instance, is an example of one type of free information source upon which NCES might rely. Among the many resources that the MMP provide are a description of multilevel models and their applications, an invitation to join a discussion group (i.e., listserv list), and references to recent articles that use multilevel models.

Strategy 4: Create Electronic Discussion Groups (or Listserv Lists)

It seems sensible for NCES to use the Internet to connect through discussion groups or listserv lists those who share common interests (e.g., licensed data users of SASS) or constitute particular technical review panels (e.g., the SASS data user's group). Whatever communications transpire among members of such data analysis groups might then be made available to those who design NCES surveys. Providing the designers with this information exploits the capability of electronic mailing list servers to permanently record all messages. There is precedent for creating discussion groups at NCES as well. The Advisory Council on Education Statistics (ACES), for example, makes use of a listserv, one form of a discussion group.

Strategy 5: Most Frequently Asked Questions and Relevant Literature on the Web

NCES data users may frequently ask NCES numerous questions about the data and its analysis. Posting these questions (and their answers) on the Web then seems sensible inasmuch as it may prevent those who respond to the questions from repeating themselves incessantly. The strategy complements NCES's emplaced effort to provide instruction to researchers who aspire to analyze NCES data. NCES, for example, holds seminars during the summers "to provide young scholars and researchers with opportunities to gain access" to NCES surveys (NCES 1994). Knowledge of the types of questions that data users frequently ask, moreover, might prove useful to those who design NCES surveys. For example, if a preponderance of questions were to pertain to the techniques required to model the measurement error that results from NAEP's use of plausible values, then survey designers might want to consider a variety of options for the design of future assessments, including use of a different method of estimating proficiency.

At times, NCES might also post entire documents "to make things easier for interested parties in terms of their hunt for relevant literature" (Maline 1993, p. iii) It may be, for instance, that data analysts frequently request a particular document, say, the annotated bibliography of NLS-72 studies. To accommodate such interested parties, NCES might convert this document either to an HTML or .pdf file, then make it available through the Web.

Strategy 6: Consider Using Adobe Acrobat to Disseminate Information

Posting Adobe's portable document files (.pdf) on the Web is a particularly attractive alternative for organizations that wish to disseminate information through the Web but resist the intensive editing that HTML requires. The US General Accounting Office, for example, makes available .pdf files for dissemination via the Web. NCES might do the same. If the original NCES document were a WordPerfect of Microsoft Word file that included several graphical figures (e.g., a data user's manual), NCES, after purchasing the reasonably priced Adobe Acrobat, would only have to issue a print command to create a .pdf file. The software has additional features that NCES might find attractive, as well. It is possible, for example, to include hypertext links (e.g., from the table of contents to the conclusion) within .pdf documents. Future versions of the software, moreover, will enable authors to include hypertext links from within .pdf documents to other Web locations (e.g., NCES's home page). Finally the Adobe Acrobat Reader, the application required to read .pdf files, operates almost seamlessly with Web browsers, particularly Netscape, and is available through the Internet at no cost for many operating systems.

Implementing the Strategies: How Difficult Is It?

Making judgments about which strategies to implement and in what order boils down to a cost-benefit analysis that NCES will have to do. What we can provide, however, are some final thoughts. We base our thoughts in large part on the effort required to create this document's illustrative Web home page at HTTP://www.dolphin.upenn.edu/~terhania/.

Strategy 1: Elicit Information Through Fill-Out Forms and Electronic Mail

Setting up a Web server to receive, process, and store information (e.g., responses to questionnaires) is straightforward, although it does require some tinkering on the server end (e.g., see Magid, Matthews, and Jones 1995). The necessary resources are in place, however, because NCES has already begun to use the Web.

Developing Web pages through HTML requires some expertise. Nevertheless, it is fairly easy to capitalize on the work of others. The HTML code that underlies the creation of each file posted on the Web is available at no cost; that is, prototypical Web pages are readily available.

Strategy 2: Distribute Spreadsheet Files Through the Web

Someone at NCES must do the work. It is possible to capitalize on the work of others, however. This is one object of creating Web windows.

Strategy 3: Track the Emergence and Development of New Analytic Methods and Their Implications Through the Web

A template for acquiring and consolidating such information might be based on the list of "implication" categories given earlier in this report. The categories, put into question form, are: What are the implications of the new analysis method or its application for deciding

1) What variables to measure;
2) How to measure;
3) Whom to measure;
4) How many respondents to sample;
5) When to measure;
6) With what sample design characteristics;
7) In connection with what other data collection;
8) Why; and
9) With what reporting strategy (e.g., CD-ROM, and so on).

Strategy 4: Create Electronic Discussion Groups (or Listserv Lists)

NCES has a list of all licensed data users. Membership of NCES's technical review panels, moreover, is public information. Further, there is precedent for using mailing list servers to connect members of particular panels or groups. Creating additional discussion groups, therefore, is a logical next step. The listserv of the Advisory Council on Education Statistics is a prototype.

Strategy 5: Post Frequently Asked Questions and Relevant Literature on the Web

Posting those questions that data users frequently ask and relevant literature on the Web requires some effort. Nevertheless, many may benefit through the work of few. Moreover, NCES might capitalize on the work of others here as well. One question that NCES and its contractors may frequently field, for example, relates to the appropriate statistical procedures that must be applied to obtain accurate variance estimates with NCES surveys, e.g., SASS. These, we presume, are the "reasonably tractable procedures" to which Clogg (1989) refers. We know, for example, that SASS analysts use, among other software, a package developed at Westat called WesVarPC to apply the procedures. We know, as well, that Westat provides documentation that includes an introduction to replication methods in portable document format (.pdf). NCES, with Westat's permission, might make this file available to analysts.

Strategy 6: Consider Using Adobe Acrobat to Disseminate Information

Using Adobe's portable document format (.pdf) does not force NCES to decide among software packages. Creating a .pdf file is as simple as issuing a print command.

NOTES

1. All data from the National Center for Education Statistics that are used here apply only to public schools (and public school students).

2. "Significantly," as used here and throughout the paper, refers to a mean difference of at least two standard deviations.

3. Alternatively, the analyst might simply use the propensity score as a covariate in an analysis of covariance—e.g., see Rosenbaum and Rubin (1983).

REFERENCES

American Statistical Association. 1993. *Proceedings of the American Statistical Association: Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Bailar, B. A., and Lanphier, C. M. 1978. *Development of Survey Methods to Assist Survey Practices*. Washington, D.C.: American Statistical Association.

Barnes, R. E., and Ginsburg, A. L. 1979. "Relevance of the RMC Models for Title I Policy Concerns." *Educational Evaluation and Policy* 1 (2): 7-14.

Barro, S. M. 1992. "Models for Projecting Teacher Supply, Demand, and Quality: An Assessment of the State of the Art." In E. E. Boe and D. M. Gilford (Eds.) *Teacher Supply, Demand, and Quality*. Washington, D.C.: National Academy Press, 129-209.

Berk, R. A. et al. 1985. "Social Policy Experimentation." *Evaluation Review* 9: 387-429.

Bernard, H. R., and Killworth, P. D. 1973. "On the Social Structure of an Ocean-Going Research Vessel and Other Important Things." *Social Science Research* 2: 145-184.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1987. "Estimating the Number of People in an Average Personal Network and an Event Subpopulation." *Proceedings of the American Statistical Association: Survey Research Methods Section*. Washington, D.C.: American Statistical Association, 17-25.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1989. "Estimating the Size of an Average Personal Network and of an Event Subpopulation." In M. Kochen (Ed.) *The Small World*. Norwood, N.J.: Ablex.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., McCarty, C., Shelly, G. A., and Robinson, S. 1990. "Comparing Four Different Methods for Measuring Personal Social Networks." *Social Networks* 12 (3): 179-215.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1990. "Estimating the Size of An Average Personal Network and of An Event Subpopulation: Some Empirical Results." *Social Science Research* 20: 109-121.

Bernard, H. R., Killworth, P. D., Johnsen, E. C., Shelley, G. A., and McCarty, C. 1994. *Estimating the Size of Uncountable Populations: A Summary of Research*. Gainsville, Florida: University of Florida.

Bernard, H. R., Johnsen, E. C., Killworth, P. D. and Robinson, S. 1994. *How Many People Died in the Mexico City Earthquake?* Gainsville, FL: University of Florida.

Betsey, C., Hollister, R., and Papgiorgiou, M. (Eds.). 1985. *The YEDPA Years: Report of the Committee on Youth Employment Programs*. Washington, D.C.: National Research Council.

Bickman, L. 1985. "Improving Established Statewide Programs." *Evaluation Review* 9: 189-208.

Bloom, H. S., Borus, M. E., and Orr, L. L. 1987. *Using Random Assignment to Evaluate an Ongoing Program: The National JTPA Evaluation*. Presented at the Annual Meeting of the Statistical Association, San Francisco, August 17-20.

Blumstein, A., Cohen, J., Roth, J., and Visher, C. A. (Eds.). 1986. *Criminal Careers and "Career Criminals."* Washington, D.C.: National Academy Press.

Board on Children and Families and Committee on National Statistics. 1995. *Integrating Federal Statistics on Children*. Washington, D.C.: National Academy Press.

Bock, D., Gibson, R., and Muraki, E. 1988. "Full Information Item Factor Analysis." *Applied Psychological Measurement* 12: 261-280.

Boe, E. E., and Gilford, D. M. (Eds.). 1992. *Teacher Supply, Demand, and Quality: Policy Issues, Models, and Databases*. Washington, D.C.: National Academy of Sciences Press.

Boruch, R. F. 1975. "Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation." *Social Methods and Research* 4: 31-53.

Boruch, R. F. 1994. "The Future of Controlled Experiments." *Evaluation Practice* 15 (3): 265-274.

Boruch, R. F. 1995. "Comments on Droitcour and Chelimsky's *Cross-Design Synthesis*." Presented at the Annual Meeting of the American Evaluation Association and the Canadian Evaluation Association, Vancouver, B.C. (Author: University of Pennsylvania, Philadelphia, PA).

Boruch, R. F., and Riecken, H. W. (Eds.). 1975. *Experimental Testing of Public Policy*. Boulder, CO: Westview.

Boruch, R. F., and Cecil, J. S. 1979. *Assuring Confidentiality of Social Research Data*. Philadelphia, PA: University of Pennsylvania Press.

Boruch, R. F., McSweeny, A. J., and Soderstrom, J. 1978. "Bibliography: Illustrative Randomized Experiments." *Evaluation Quarterly*.

Boruch, R. F., and Wothke, W. 1985. "Seven Kinds of Randomization Plans for Designing Field Experiments." *New Directions for Program Evaluation* 28: 95-118. San Francisco: Jossey-Bass.

Boruch, R. F., and Pearson, R. W. 1988. "Assessing the Quality of Longitudinal Surveys." *Evaluation Review* 12 (1): 3-59.

Bradburn, N., and Sudman, S. 1981. *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

Breslow, N. 1989. "Biostatistics and Bayes." In M. Gail and N. Johnson, Coordinators (Sesquicentennial Invited Paper Sessions: Proceedings of the American Statistical Association.) Alexandria, VA: American Statistical Association, 51-69.

Breslow, N. E., and Clayton, D. G. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88: 9-25.

Brooks-Gunn, J., Brown, B., Duncan, B., and Moore, K. A. 1995. "Child Development in the Context of Family and Community Resources." In Board on Children and Families. *Integrating Federal Statistics on Children: Report on a Workshop*. Washington, D.C.: National Academy Press, 27-97.

Broward County School Board. Department of Research. 1987. *Achievement Through Instruction and Motivation: Program Evaluation Report for 1986-87*. Fort Lauderdale, FL: School Board of Broward County, Research Department.

Bruce, R. V. 1973. *Bell: Alexander Graham Bell and the Conquest of Solitude*. New York, NY: Little Brown.

Bryk, A. S., and Raudenbush, S. W. 1992. *Hierarchical Linear Models: Applications and Data Analysis*. Newbury Park, CA: Sage.

Bryk, A., Raudenbush, S., Seltzer, M. and Conger, R. 1989. *An Introduction to HLM: Computer Program and User's Guide*. Chicago: University of Chicago, Department of Education.

Campbell, D. T., and Boruch, R. F. 1975. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives." In C. A. Bennett and A. A. Lumsdaine (Eds.) *Central Issues in Social Program Evaluation*. New York: Academic Press, 195-297.

Chelimsky, E. (Ed.). 1985. *Program Evaluation: Patterns and Directions*. Washington, D.C.: American Society for Public Administration (PAR Classics Series).

Coleman, J. S., Hoffer, T., and Kilgore, S. 1982. *High School Achievement: Public, Catholic, and Private Schools Compared*. New York: Basic Books.

Clogg, C. C. 1989. "Modeling Social Statistics: Current Issues." In M. H. Gail and N. L. Johnson, Coordinators. (Sesquicentennial Invited Paper Sessions: Proceedings of the American Statistical Association.) Alexandria, VA: American Statistical Association, 214-225.

Cohen, M. 1994. "Intergrated Sampling of Education Institutions." (Proceedings of the American Statistical Association Survey Research Methods Section.) Alexandria, VA: American Statistical Association, 638-640.

Cooley, W. W. 1988. *Design for a Longitudinal Study of Chapter I.* Briefing to the U.S. Department of Education, Washington, D.C.

Cottingham, P., and Rodriguez, A. 1987. *The Experimental Testing of the Minority Female Single Parents Program.* Presented at the Annual Meeting of the American Statistical Association, San Francisco, CA, August 17-20.

Coyle, S., Boruch, R. F., and Turner, C. F. (Eds.). 1991. *Evaluating AIDS Prevention Programs.* Washington, D.C. National Academy Press.

Cox, L. H., and Boruch, R. F. 1988. "Record Linkage, Privacy, and Statistical Policy." *Journal of Official Statistics* 4 (1): 3-16.

Cronbach, L. J. et al. 1980. *Toward Reform of Program Evaluation.* San Francisco, CA: Jossey-Bass.

Davis, C., and Sonnenberg, B. (Eds.). 1993. *Programs and Plans of the National Center for Education Statistics: 1993 Edition.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Davis, C. and Sonnenberg, B. (Eds.). 1995. *Programs and Plans of the National Center for Education Statistics: 1995 Edition.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Draper, D. et al. 1992. *Combining Information for Research.* Washington, D.C.: National Academy of Sciences. (Also in: *Contemporary Statistics.* #1, Alexandria, VA: American Statistical Association, Undated).

Draper, D. 1995. "Inference and Hierarchical Modeling in the Social Sciences." *Journal of Educational and Behavioral Statistics* 20 (2): 115-148.

Droitcour, J., and Chelimsky, E. 1995. *Cross-Design Synthesis.* Paper presented at the Annual Meeting of the American Evaluation Association and the Canadian Evaluation Association, Vancouver, B.C. (Authors: U.S. General Accounting Office, Program Evaluation and Methodology Division, Washington, D.C.)

Droitcour, J. A., Silberman, G., and Chelimsky, E. 1993. "Design Synthesis." *International Journal of Technology Assessment in Healthcare* 9 (3): 440-449.

Duncan, G. J., and G. Kalton. 1985. *Issues of Design and Analysis of Surveys Across Time.* Presented at the centenary session of the International Statistical Institute, Amsterdam.

Duncan, G. J., Juster, F. T., and Morgan, J. N. 1984. "The Role of Panel Studies in a World of Scarce Research Resources." In S. Sudman and M.A. Spaeth (Eds.) *The Collection and Analysis of Economic and Consumer Behavior Data: In Memory of Robert Ferber.* Champaign, IL: Bureau of Economics and Business Research.

Elmore, R.F. 1993. "What Knowledge Base?" *Review of Educational Research* 63: 314-318.

Farrington, D. P. 1988. "Advancing Knowledge About Delinquency and Crime: The Need for a Coordinated Program of Longitudinal Research." *Behavioral Sciences and Law* 6 (3): 307-331.

Farrington, D. P., Ohlin, L. E., and Wilson, J. Q. 1986. *Understanding and Controlling Crime: Toward a New Research Strategy*. New York: Springer-Verlag.

Fienberg, S. B., and Tanur, J. 1986. "From the Inside Out and the Outside In: Combining Experimental and Sampling Structures." *Technical Report 373*, Carnegie-Mellon University (December).

Fienberg, S. B., and Tanur, J. 1987a. "The Design and Analysis of Longitudinal Surveys: Controversies and Issues of Cost and Continuity." In R. F. Boruch and R. W. Pearson (Eds.) *Designing Research with Scarce Resources*. New York: Springer-Verlag, 60-93.

Fienberg, S. B., and Tanur, J. 1987b. "Experimental and Sampling Structures: Parallels Diverging and Meeting." *International Statistics Review* 55: 75-96.

Fienberg, S. B., Singer, B., and Tanur, J. 1985. "Large Scale Social Experimentation in the United States." In A. C. Atkinson and S. E. Fienberg (Eds.) *A Celebration of Statistics: The ISI Centenary Volume*. New York: Springer-Verlag, 287-326.

Fienberg, S. B., Martin, M. E., and Straf, M. L. 1985. *Sharing Research Data*. Washington, D.C.: National Academy of Sciences.

Folsom, R. E. and Liu, J. 1994. "Small Area Estimator for the National Household Survey of Drug Abuse." *Proceedings of the Section on Survey Research Methods: Annual Meeting of the American Statistical Association*. Alexandria, VA: ASA, 565-570.

Fraker, T., and Maynard, R. 1987. *The Use of Comparison Group Designs in Evaluation of Employment Related Programs*. Princeton, NJ: Mathematica Policy Research.

Fraker, T., and Maynard, R. 1987. "The Use of Comparison Group Designs for Evaluations of Employment-Related Programs." *The Journal of Human Resources* 22: 194-227.

Frederikson, C. H., and Rotondo, J. A. 1979. "Time Series Models and the Study of Longitudinal Change." In J. R. Nesselroade and P. B. Baltes (Eds.) *Longitudinal Research in the Study of Behavior and Development*, pp. 111–154. New York: Academic Press.

Freedman, D. A. 1985. "Statistics and the Scientific Method." In W. M. Mason and S. E. Fienberg (Eds.) *Cohort Analysis in Social Research*, pp. 343–36. New York: Springer-Verlag, 334-36.

Gail, M. H., and Johnson, N. L. (Eds.) 1989. Sesquicentennial Invited Paper Sessions Proceedings of the American Statistical Association. Alexandria, VA: ASA.

Gerald, D., and Hussar, W. J. 1992. *Projections of Education Statistics to 2000*. Washington, D.C.: National Center for Educational Statistics (NCES 92-218).

Gray-Donald, K., and Kramer, M. S. 1988. "Causality Inference in Observations Versus Experimental Studies." *American Journal of Epidemiology* 127: 885-892.

Greenwood, P. July 1988. *The Role of Planned Interventions in Studying the Assistance of Criminal Behavior in a Longitudinal Study*. Concept Paper Developed for the Resistance Group of the Program on Human Development. Santa Monica: RAND Corporation.

Griffith, J. 1992. Presentation to the National Advisory Council on Education Statistics March 12-13, 1992: *Draft Paper on a Proposal for an Integrated Longitudinal Studies Program*. Washington, D.C.: National Center for Education Statistics.

Gueron, J. M. 1985. "The Demonstration of State Work/Welfare Initiatives." *New Directions for Program Evaluation* 28: 5-13.

Hamburg, B. 1993. *New Futures for the Forgotten Half: Realizing Unused Potential for Learning and Productivity: William T. Grant Foundation Annual Report*. New York: The William T. Grant Foundation.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I .M., and Snow, R. E. 1995. "Enhancing the Validity and Usefulness of Large Scale Assessments: II. NELS:88. Science Achievement." *American Educational Research Journal* 32 (3): 555-582.

Heckman, J., and Singer, B. (Eds.). 1985. *Longitudinal Analysis of Labor Market Data*. Chicago, IL: University of Chicago Press.

Heckman, J. J., and Robb, Jr., R. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In J. J. Heckman and B. Singer (Eds.) *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press, 156-246.

Hedges, L. V., and Nowell, A. 1995. "Sex Differences in Mental Test Scores, Variability,and Numbers of High Scoring Individuals." *Science* 269: 41-45.

Hedges, L. V., and Olkin, I. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Heubert, J. 1992. *Personal Communication: Class Notes from Course in Law and Education*. Cambridge, MA: Graduate School of Education, Harvard University.

Hilton, T. (Ed.). 1992. *Using National Databases in Educational Research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoffer, T. B. 1992. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." *Educational Evaluation and Policy Analysis* 14 (3): 205-227.

Hoffreth, S. 1995. "Children's Transition to School." In Board on Children and Families and Committee on National Statistics. *Integrating Federal Statistics on Children: Report of A Workshop*. Washington, D.C.: National Academy Press, 98-121.

Holland, P. W., and Rubin, D. B. (Eds.). 1982. *Test Equating*. New York: Academic.

Howard, K. et al. 1988. *A Survey of Adolescents and Their Access to Mental Health Services*. Evanston, IL: Psychology Department, Northwestern University.

Howe, D. 1995. *The Free On-Line Dictionary*. (HTTP://wombat.doc.ic.ac.uk/foldoc?Free+On-line+Dictionary).

Hunter, J. E., and Schmidt, F. L. 1990. *Methods of Meta-Analysis*. Newbury Park, CA.: Sage Publications.

Johnsen, E. C., Bernard, H. R., Killworth, P. D., Shelley, G. A., and McCarty, C. 1994. "A Social Network Approach to Corroborating the Number of AIDS/HIV+ Victims in the U.S." *NERC Oceanography Unit*. Parks Road, Oxford, England: Clarendon Laboratory.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., and McCarty, C. 1990. "Estimating the Size of Personal Networks." *Social Networks* 12 (4): 289-312.

Killworth, P. D., McCarty, C., Johnsen, E. C., Shelly, G. A., and Bernard, H. R. 1994. "A Social Network Approach to Estimating Seroprevalence in the United States." *NERC Oceanography Unit*. Parks Road, Oxford, England: Clarendon Laboratory.

Killworth, P. D. et al. Undated. "Estimation of Seroprevalence, Rape, and Homelessness in the U.S. Using a Social Network Approach." *NERC Oceanography Unit*. Parks Road, Oxford, England: Clarendon Laboratory.

Kilss, W., and Alvey, W. (Eds.). 1985. *Record Linkage Techniques: Proceedings of the Workshop on Exact Matching Methodologies*. Washington, D.C.: U.S. Department of Treasury, Statistics of Income Division, IRS.

Kreft, I. G. (Ed.). 1995. "Hierarchical Models." *Journal of Educational and Behavioral Statistics* 20 (22): 109-240.

Kruskal, W. H. (Ed.). 1982. *The Social Sciences: Their Nature and Use*. Chicago: University of Chicago Press.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., and Snow, R. E. 1995. "Enhancing the Validity and Usefulness of Large Scale Educational Assessments: 1. NELS:88 Mathematics Achievement." *American Educational Research Journal* 32 (3): 525-554.

LaLonde, R. 1986. "Evaluating the Econometrics Evaluations of Training Programs with Experiments." *American Economic Review* 76 (4): 604-620.

Laumann, E. O., Gagnon, J. H., Michaels, M. S., Michael, R. T., and Coleman, J. S. 1989. "Monitoring the AIDS Epidemic in the United States: A Network Approach." *Science* 244: 1186-1189.

Lee, V. E. 1992. "Pooling Data from Two Longitudinal Cohorts." In T.L. Hilton (Ed.) *Using National Data Bases in Educational Research*. Hillsdale, NJ: Lawrence Erlbaum, 246-258.

Linn, R. L. 1979. "Validity of Inferences Based on the Proposed Title I Evaluation Models." *Educational Evaluation and Policy Analysis* 1 (2): 23-32.

Little, R. J. A., and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York, NY: Wiley.

Magid, J., Matthews, R. D., and Jones, P. 1995. *The Web Server Book: Tools & Techniques for Building Your Own Information Site*. Chapel Hill, NC: Ventana Press.

Malec, D. 1993. Chapter 8. "Model Based State Estimates from the National Health Interview Survey." In Subcommittee on Small Area Estimation, Federal Committee on Statistical Policy." *Statistical Policy Working Paper 21. Indirect Estimators in Federal Programs*. Washington, D.C.: U.S. Office of Management and Budget.

Maline, M. S. 1993. *The National Longitudinal Study of the High School Class of 1972: Annotated Bibliography of Studies*. Washington, D.C.: Office of Research, U.S. Department of Education.

Malvin, J. H., and Moskowitz, J. M. 1983. "Anonymous Versus Identifiable Self Reports of Adolescent Drug Attitudes, Intention and Use." *Public Opinion Ouarterly* 47: 557-566.

Manski, C. R. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mason, W. M., and Fienberg, S. E. (Eds.). 1985. *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York: Springer-Verlag.

Mathiowetz, N. A., and Duncan, G. J. 1984. "Temporal Patterns of Response Errors on Retrospective Reports of Unemployment and Occupation." *Proceedings of the American Statistical Association: Section Q Survey Research Methods*, pp. 652–654. Washington, D.C.: American Statistical Association.

Mathiowetz, N. A. 1987. "Response Error: Correlation Between Estimation and Episodic Recall Tasks." *Proceedings of the American Statistical Association: Survey Research Methods Section*, pp. 430–435. Washington, D.C.: American Statistical Association.

Maynard, R. A. 1987. "The Role of Randomized Experiments in Employment Training Evaluations." *Proceedings of the American Statistical Association: Survey Research Methods Section*, pp. 109–113. Washington, D.C.: American Statistical Association.

Mazur, A., and Boyko, E. 1981. "Large-Scale Ocean Research Projects: What Makes Them Succeed or Fail?" *Social Studies of Science* 11: 425-449.

Messick, S. 1984. "A New Design for the National Assessment of Education Progress." *Proceedings of the American Statistical Association: Survey Research Methods Section*. Washington, D.C.: American Statistical Association.

McCarty, C. et al. 1995. *Eliciting Representative Samples of Personal Network*. Gainsville, FL: University of Florida.

Mok, M. June 1995. "Sample Size Requirements for 2-Level Designs in Educational Research." *Multilevel Modeling Newsletter*. June 1995.

Mosteller, F., Light, R., and Sachs, J. 1995. "Sustained Inquiry in Education: Lessons from Ability Grouping and Class Size." *Center for Evaluation of the Program on Initiatives for Children*. Cambridge, MA: Harvard University.

Mosteller, F., and Moynihan, D. (Eds.). 1972. *On Equality of Educational Opportunity*. New York: Vintage Books.

Mullis, I. V. S., Jenkins, F., and Johnson, E. G. 1994. *Effective Schools in Mathematics: Perspectives from the 1992 NAEP Assessment*. Research and Development Report. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Mundel, D. 1979. "Memo to Franklin Zweig" (November 15). Congressional Budget Office.

Murnane, R. J. 1992. "Who Will Teach?" In E. E. Boe and D. M. Gilford (Eds.) *Teacher Supply, Demand, and Quality*, pp. 262–270. Washington, D.C.: National Academy Press.

National Center for Education Statistics. 1995. *Agenda for a Meeting on the Future of, Education Statistics*. Berkeley, CA: MPR Associates.

National Center for Education Statistics. 1994. *Announcement: Advanced Studies Seminar on the Use of NELS:88 and SASS Data for Research and Policy Discussion*. Washington, DC: NCES, U.S. Department of Education.

National Center for Education Statistics. 1993. *National Adult Literacy Survey*. Washington, D.C.: NCES.

Netscape Communications Corporation. 1995. *White paper*. (HTTP://home.netscape.com/comprod/at_work/white_paper/index.HTML).

Oakes, J. 1990. *Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Sciences*. Santa Monica, CA: RAND.

Office of Management and Budget. 1993. *Statistical Policy Paper 21. Indirect Estimators in Federal Programs*. Washington D.C.: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget (Subcommittee on Small Area Estimation/Federal Committee on Statistical Metrology).

Ohlin, L. E. May 12, 1988. "Memo to Working Group on Desistance Regarding Policy Statement as Program Objective." *Program on Human Development and Criminal Behavior*. Castine, Maine.

Osgood, D. W., and Smith, E. L. 1995. "Applying Hierarchical Linear Modeling to Extended Longitudinal Surveys." *Evaluation Review* 19 (1): 3-30.

Pallas, A. 1995. "Federal Data on Educational Attainment and the Transition to Work." In Board on Children and Families and Committee on Federal Statistics. *Integrating Federal Statistics on Children: Proceedings of a Workshop*, pp. 122–155. Washington, D.C.: National Academy Press.

Pearson, R. W. 1987. *Researchers' Access to U.S. Federal Statistics*. Items 41: 6-11.

Pearson, R. F. and Boruch, R. F. (Eds.). 1986. *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits*. (Lecture Notes in Statistics, N. 38.) New York: Springer-Verlag.

Project Review Team. 1988. "Report on the Spouse Assault Replication Project to the National Institute of Justice." *Department of Statistics and Psychology*. Northwestern University, Evanston, IL.

Raish, M. 1994. *Network Knowledge for the Neophyte*. Binghamton, NY: Binghamton University Libraries.

Reisner, E. R., Alkin, M. C., Boruch, R. F., Linn, R. L., and Millman, J. 1982. *Assessment of the Title I Evaluation and Reporting System*. Washington, D.C.: U.S. Department of Education.

Reisner, E. R., Turnbull, B. J., and David, J. L. 1988. *Evaluation of the ECIA Chapter I Technical Assistance Centers*. Washington, D.C.: Policy Studies Associates, Inc.

Reiss, A. J. 1986. "Co-Offending Influences on Criminal Careers." In A. Blumstein, J. Cohen, R. Roth, and C. Visher (Eds.) *Criminal Careers and "Criminal Careers" Volume I*. Washington, D.C.: National Academy of Sciences.

Reiss, A. et al. 1988. "Pipeline Studies in the Spouse Assault Replication Project." In Report of the Program Review Team, Spouse Assault Replication Project, to the National Institute of Justice. Departments of Statistics and Psychology. Northwestern University, Evanston, IL.

Riecken, H. W. et al. 1974. *Social Experimentation*. New York: Academic Press.

Rock, D. A. 1992. "Pooling Results from Two Cohorts Taking Similar Tests." In T.L. Hilton (Ed.) *Using National Data Bases in Educational Research*. Hillsdale, NJ: Lawrence Erlbaum, 192-213.

Rogosa, D., and Saner, H. 1995. "Longitudinal Data Analysis Examples with Random Coefficient Models." *Journal of Educational and Behavioral Statistics* 20 (2): 149-170.

Rosen, S. (Ed.). 1974. *Final Report of the Panel on Manpower Training Evaluation: The Use of Social Security Earnings Data for Assessing the Impact of Manpower Training Programs*. Washington, D.C.: National Academy of Sciences.

Rosenbaum, P. R. 1986. "Dropping Out of High School in the United States: An Observational Study." *Journal of Educational Statistics* 11 (3): 207-224.

Rosenbaum, P. R. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science* 2 (3): 92-316.

Rosenbaum, P. R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 104-1032.

Rosenbaum, P. R. 1991. "A Characterization of Optimal Designs for Observational Studies,"*Journal of the Royal Statistical Society* B53 (3): 597-610.

Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika* 70 (1): 41-55.

Rosenbaum, P. R., and Rubin, D. B. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516-524.

Rosenbaum, P. R., and Rubin, D. B. 1982. "Comparing Effect Sizes of Independent Studies." *Psychological Bulletin* 92: 500-504.

Rubin, D. B. 1974. "Estimating Causal Effect of Treatment In Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688-701.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Scheuren, F. 1985. "Methodological Issues in Linkage of Multiple Databases." In B. Kilss and W. Alvey (Eds.) *Record Linkage Techniques*. Washington, D.C.: U. S. Department of Treasury, Statistics of Income Division, Internal Revenue Service.

Scheuren, F. 1985. "Methodologic Issues in Linkage of Multiple Databases." Prepared for the Panel on Statistics for the Aged Population, National Academy of Sciences. Washington, D.C.: National Academy of Sciences.

Scheuren, F. 1995. *Administrative Record Opportunities in Educational Survey Research.* Report prepared for the National Center on Educational Statistics. Washington, D.C.: The George Washington University.

Shaffer, J. P. (Ed.). 1992. "The Role of Models in Nonexperimental Social Science: Two Debates." *Journal of Educational Statistics* (Special Issue).

Slavin, R. E. 1993. "Ability Grouping in the Middle Grades: Achievement Effects and Alternatives." *Elementary School Journal* 93 (5): 535-552.

Smith, M. 1988. *Thoughts on the Chapter I Longitudinal Evaluation Design*, Briefing to the U.S. Department of Education. Washington, D.C.

Spencer, B. D. 1980. *Benefit-Cost to Allocate Funds*. New York: Springer-Verlag.

St. Pierre, R., Schwartz, J., Murray, S., Deck, D., and Nickel, P. 1993. *National Evaluation of Even Start Family Literacy Program* (Contract 9006-2001). Washington, D.C.: U.S. Department of Education.

Stafford, F. 1985. "Forestalling the Demise of Empirical Economics: The Role of Microdata in Labor Economics Research." In O. Ahsenfelter and R. Layard (Eds.) *Handbook of Labor Economics*. New York: North-Holland.

Stone, E. F., Gardner, D. G., Gueutal, H. G., and McClure, S. 1983. "A Field Experiment Comparing Information Privacy Values, Beliefs, and Attitudes Across Several Types of Organizations." *Journal of Applied Psychology* 68: 459-468.

Taeuber, R., and Rockwell, R. C. 1982. "National Social Data Series: A Compendium of Brief Descriptions." *Review of Public Data Use* 10: 23-111.

U.S. Department of Education, National Center for Education Statistics. 1993 *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and States.*

U.S. General Accounting Office. 1986. *Computer Matching: Assessing Its Costs and Benefits* (PEMD-87-2) Washington, D.C.: USGAO.

U.S. General Accounting Office. 1986. *Computer Matching: Factors Influencing the Agency Decision Making Process* (PEMD-87-3 BR) Washington, D.C.: USGAO.

U.S. General Accounting Office. 1992. *Cross-Design Synthesis: A New Strategy for Medical Effectiveness Research* (GAO/PEMD-92-18) Washington, D.C.: USGAO.

U.S. General Accounting Office 1995 *Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies* (GAO/PEMD-95-9) Washington, D.C.: USGAO.

Verdonik, F., and Sherrod, L. R. 1984. *An Inventory of Longitudinal Research on Childhood and Adolescence*. New York: Social Science Research Council.

WebCrawler. 1995. *WebCrawler News*. (HTTP://Webcrawler.com/WebCrawler/Facts/Size.HTML).

Weis, J. G. 1987. "Issues In the Measurement of Criminal Careers." In A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher (Eds.) *Criminal Careers and "Career Criminals*." Washington, D.C.: National Academy Press, 1986.

Weiss, A. 1988. *Randomized Experiments and Time Series Analysis in Police Research*. Evanston, IL: Department of Political Science, Northwestern University.

White, S. (Ed.). 1994. *Overview of NAEP Assessment Frameworks*. Washington, D.C.: NCES, U.S. Department of Education National Center for Education Statistics.

Wilson, D., Wood, R., and Gibbons, R. D. 1991. *ESTFACT: Test Scoring, Item Statistics,and Item Factor Analysis*. Chicago, IL: Scientific Software.

Zellner, A. 1989. "Discussion." In Mitchell H. Gail and N.L. Johnson (Coords.) *Proceedings ofthe American Statistical Association, Sesquicentennial Invited Papers*. Alexandria: American Statistical Association, 162-166.

Discussant Comments

FREDERICK MOSTELLER

I am extremely impressed with the paper by Bob Boruch and George Terhanian, partly because in several instances they address matters totally new to me. I anticipate that their paper will repay study by the staff of NCES for a long time. I shall comment on only a few of the many issues they treat.

What makes their paper especially effective is the way it appreciates problems of methodology as well as substance and, as Emerson Elliott recommended in his opening speech, how it blurs the distinction between statistics and research and between retrieval and dissemination. Their ability to make connections between different fields and to suggest enterprises that have interest for multiple agencies enhances the opportunities to serve the public by informing Americans about the state of various problems in education. And Boruch and Terhanian have also a beneficial, insightful capacity to see what sorts of activities will engage the attention of an administration, a Congress, or a public proceeding down the information highway.

For example, encouraging people to interpret their own analytic contributions and those of others in order to improve the design of sample surveys is certainly good advice. I had not formerly thought about such a move.

Droitcour and Silberman of the GAO have given us a great challenge in developing the idea of cross-design synthesis. It is especially appropriate to think of its possibilities as a way of using sample surveys to strengthen inferences from experimentation. Their general idea is to let weaknesses from one form of investigation be buttressed by strength from another method, for example, by balancing biases. This good idea needs extensive development.

In order to achieve this goal, we need many investigators to carry out practical examples. From a collection of such examples, we may be able to sieve out principles that can be used in other circumstances. So far, we do not have many examples.

With so few examples of applications, we cannot yet speak of cross-design synthesis as a working method, but when the examples grow in number, we will have a new technique. It will be useful to have NCES encouraging the use of sample surveys to broaden the variety of devices available for cross-design synthesis.

NCES can also develop the ideas mentioned by Boruch and Terhanian that would additionally help link ideas between experiments and surveys. This requires knowing what experiments are being carried out and which ones might be usefully linked to one another by suitable future surveys. For example, can local experimental treatments increase performance in a region? Can surveys measure improvements in a region flowing from local programs of education or of health, such as disease prevention?

Boruch and Terhanian discuss what I like to call "skill grouping," rather than "ability grouping," where classes with different skill levels are put into homogeneous classes rather than heterogeneous ones. Presumably the hope is that students in homogeneous classes will learn more than those in heterogeneous groupings. They present NAEP data that implies that students in homogeneous classes perform better in 8th grade arithmetic than those in heterogeneous classes in most jurisdictions in 1990 and 1992.

In the 10 randomized (or nearly randomized) experiments my colleagues found to review (not restricted to arithmetic), the average performance over different subjects was about equal for the skill-grouped (homogeneous) and the whole-class (heterogeneous) instruction. Although the variation in outcome from study to study was substantial, the reporting was often inadequate. There was also no real way to appreciate whether the students in the studies represented the nation in any reasonable sense. Moreover, no experiment lasted more than 1 year (at least in the most popular form of skill grouping), and each experiment represented only one school.

The contrast between the outcomes in the sample survey and the experiment deserves more explanation. This is an example of an issue whose study might be aided by a compilation by *topic of investigation* of experiments, surveys, and demonstration programs. I do not mean to include analysis, however, which is merely a map of the territory. Most investigators would like their studies included in such a list. Consequently, making such a collection may be feasible. Investigators such as myself would find such information very useful.

The only substantial educational experiment I have come across like this has been the Tennessee Class Size experiment. I have concluded that we need more such experiments.

One might hope that even though schooling is primarily run by the states that some organization could bring together groups of districts regionally or even in a national sample to carry out experiments that would have more than a single state participating. A compendium of surveys, experiments, and demonstrations might help school districts and states think of opportunities to cooperate in such ventures.

In the fourth section of their paper, Boruch and Terhanian discuss work on people who are hard to count and on measurements that are hard to make. They suggest special methods of questioning. With respect to guessing unknown numbers, I have discussed the possibility of trying to estimate the unknown numbers by independently using several different approaches. I call this process "triangulation." To accomplish this, essentially one sets up several different models, and by guessing or knowing parameters of the models, one tries to construct estimates from each model. If the models differ in structure but produce similar outcomes, this seems to give some evidence favoring the resulting estimates. I suggest that adding the idea of several approaches to these difficult measurement and counting problems may help to develop new methods of assessment.

When faced with such a plethora of suggestions as Boruch and Terhanian supply, one is tempted to try to prioritize the list. But as Elliott suggested, much of what will be feasible in the near future will depend on the *accidents* of perceived joint interests of otherwise independent organizations, and so trying to prioritize these suggestions would not be very profitable, as compared with having it done by someone who is more familiar with the current goals of the Center and its interactions with other organizations. It would be valuable to have individuals at the Center who are well prepared to work cooperatively, and this paper and others presented here certainly are making major contributions to that end.

5

New Data Collection Methodologies, Part II: Experimental Design

Incorporating Experimental Designs Into New NCES Data Collection Methodologies

---

Charles E. Metcalf

ABSTRACT

This paper considers some potential methods of accommodating policy evaluations using a formal experimental design—that is, with randomized treatment and control groups—within NCES national data collection efforts. The paper first addresses some limitations in using national data sets for selecting comparison groups for policy evaluations, and then explores the following approaches to integrating experimental designs into ongoing longitudinal databases:

- Designing a specific experiment for implementation at the initiation of the longitudinal survey, using a within-survey treatment group that receives the policy intervention;
- Designing a longitudinal survey to accommodate as-yet-unspecified future experiments;
- Augmenting a survey with supplemental sampling units that receive an experimental intervention, or expanding the longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and
- Providing a sample frame for the random selection of schools to test school-based innovations.

The paper draws the following five conclusions:

1) Because the descriptive value of NCES national data sets for framing policy issues ought not to be minimized, precautions should be taken so that efforts to accommodate policy experiments do not dilute this value.
2) Attempts to improve the attractiveness of national data sets as general-purpose (nonrandomized) comparison groups would not be warranted, because the intrinsic weakness of comparison groups relative to randomized control groups makes this effort an unpromising investment.
3) Experiments are difficult to incorporate into a national longitudinal sample, unless the timing of a demonstration implementation converges fortuitously with the initiation of a longitudinal survey that has a compatible age cohort. There is potential for improving the efficiency and comparability of longitudinal data collection, however, and for moving such experiments in the direction of using representative sample frames compatible with national data frames.

4) Attempts to append an experiment to a longitudinal survey after the survey's initiation point would be fraught with difficulties, unless supplemental samples are drawn for the demonstration treatment and control groups.

5) The potential for implementing demonstrations with across-school random assignment appears to have been severely underestimated, both for student- and school-targeted initiatives.

BACKGROUND ON THE USE OF EXPERIMENTAL DESIGNS

Since the first income maintenance experiments in the late 1960s, experimental methods that involve the random assignment of a target population to treatment and control groups have proved to be both feasible and extremely valuable for evaluating social programs and policy interventions. This approach has been established as the most defensible method for determining the extent to which *specific* policy interventions affect behavior or outcomes of interest.

Randomized experiments have been used to test interventions in such areas as welfare reform, employment and training, food stamp benefit cashout, health care delivery, long-term care, medical treatment, offender rehabilitation, domestic violence, and family preservation services. Evaluations of the Upward Bound program (funded by the U.S. Department of Education [ED]) and the Job Corps program (funded by the U.S. Department of Labor [DOL]) have broken new ground in measuring the impacts of *existing* broad-based programs by diverting nationally representative samples of program applicants into randomized control groups.[1] In addition to Upward Bound, ED has funded other recent randomized studies, including evaluations of the Dropout Demonstration Assistance Program, Dropout Prevention and Reentry projects in vocational education, the Even Start program, and Workplace Literacy programs.

While program evaluation methodology was evolving, the National Center for Education Statistics (NCES) initiated a series of longitudinal studies "to provide ongoing, descriptive information about what is occurring at the various levels of education and the major transition phases of students' lives," beginning with the National Longitudinal Study of 1972 (NLS–72) (NCES 1995). Similarly, other large-scale data sets, such as the Survey of Income and Program Participation (SIPP), the Panel Study of Income Dynamics (PSID), and the National Longitudinal Survey–Youth Cohort (NLSY) track representative samples of some of the same populations targeted by programs subject to demonstration evaluations.[2]

Yet almost invariably, program evaluations based on control group or nonrandomized comparison group methodologies have involved independent data collection efforts, usually using samples and demonstrations sites that are not nationally representative.[3] These evaluations do not take advantage of the existing array of continuing large-scale data collection efforts that might include representative samples of the potentially relevant target population of interest, except possibly for limited benchmark purposes.

Why is this? Are there deficiencies in national data sets that can be remedied, from a policy impact evaluation perspective, without compromising the primary focus of these data sets?

If program evaluations could use general-purpose databases effectively, this seemingly inefficient use of data collection resources could be rectified.

In judging the efficacy of national data sets for the evaluation of education policy, I should stress that the NCES national data sets are used for both descriptive and evaluation purposes. They are available to a wide variety of potential users as data sets for evaluating both education policy and the dynamics of educational processes and student behavior. Accurate, representative descriptions are essential for understanding an economic or policy sector: providing incontrovertible evidence of what is happening is a legitimate and primary focus of national data collection efforts that should not be compromised. In my experience, some of the greatest revelations of research projects have involved description and documentation of facts that turned out to be controversial, rather than sophisticated evaluation of policy demonstrations or experiments.

This paper considers some potential methods of accommodating policy evaluations using a formal experimental design—that is, with randomized treatment and control groups—within NCES national data collection efforts. Examples of these methods might include the following:

- Designing a specific experiment for implementation at the initiation of the longitudinal survey, using a within-survey treatment group that receives the policy intervention;
- Designing a longitudinal survey to accommodate as-yet-unspecified experiments in the future;
- Augmenting a survey with supplemental sampling units that receive an experimental intervention, or expanding the longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and
- Providing a sample frame for the random selection of schools to test school-based innovations.[4]

The next section of this paper addresses some limitations in using national data sets for selecting comparison groups for policy evaluations. The third section explores methods for adapting national data sets to accommodate formal policy experiments. The paper concludes with a brief reality assessment of approaches showing the most promise.

Using National Data Sets to Select Comparison Groups for Policy Evaluations

Longitudinal and repeated cross-sectional data sets permit many insightful analyses of causal relationships and policy impacts, but their use falls short of conventional experimental standards for measuring program impacts. They are often proposed, and sometimes used, to create comparison groups for demonstrations of a policy implemented with a separate sample of students and/or schools. But repeatedly these data sets are rejected in favor of independently collected data sets for control or comparison groups.[5]

When considering the use of an existing data set as a comparison group, designers of demonstrations and policy evaluations are confronted with a major cost advantage over the use of an independent control or comparison group and its associated data collection costs. They are also confronted with two major classes of disadvantages from a methodological standpoint. These disadvantages are associated with 1) characteristics of specific data sets relative to those of an independently tailored comparison group, and 2) general deficiencies of comparison groups relative to randomly selected control groups.

Criteria for Evaluating Existing Data Sets as Comparison Groups

Aside from general problems associated with nonrandomized comparison groups, an existing data collection vehicle would have to meet several basic requirements to be a suitable substitute for an independently defined comparison group:

- The sample must contain an identifiable subgroup that is comparable to the group receiving the demonstration treatment;
- The subsample meeting target group requirements must be large enough to meet the statistical precision requirements of the planned evaluation;
- The survey should have a longitudinal structure for tracking individual outcomes for a period comparable in length to that used for tracking the demonstration treatment group, ideally for the same period in chronological time;[6] and
- The survey database must contain comparable data elements, both for measuring background characteristics of sample members and for defining outcome measures.

To provide a concrete example of how these criteria were applied, the following describes the process by which existing longitudinal surveys were considered for use as a comparison group for the Job Corps evaluation that is currently under way. I chose this example because of my firsthand involvement in the design effort, even though Job Corps is funded by DOL rather than by ED.

The Job Corps program provides a range of education, vocational training, and support services in a predominantly residential setting to disadvantaged youths between the ages of 16 and 24.[7] Approximately 60,000 new enrollees are served each year. In 1993, DOL initiated an evaluation of the program that eventually adopted a randomized design in which approximately 8 percent of all eligible Job Corps applicants were assigned to a control group. Sample intake began in November 1994 and is scheduled to end in early 1996.

Before adopting a randomized design, we considered using an independently constructed comparison group (not discussed here) and several existing surveys—the National Education Longitudinal Study (NELS), SIPP, PSID, NLSY, and the Current Population Survey (CPS). To fulfill the criteria discussed earlier for the requirements of the Job Corps evaluation, an existing survey would have to provide a comparison group with the following characteristics:

- A representative sample of youths aged 16 to 24 in 1995, who meet specific definitions of being disadvantaged and having limited employment opportunities;

- A longitudinal structure providing outcome data for 36 to 48 months after the 1995 enrollment window; and
- Outcome measures of employment and education experience, transfer receipts, and criminal activities.[8]

None of the considered data sets could have identified eligible youths in a manner strictly comparable to the criteria applied in the Job Corps recruitment process, but all could have provided acceptable approximations of the relevant population. The NELS sample, which started as a cohort of 1988 8th graders, would have provided a sample of about 2,000 Job Corps eligibles aged 20 to 21 in 1994, but it would not have covered the full age span of eligibles. The planned 1998 NELS survey would have provided detailed education and training outcome measures 36 months after the enrollment window for the Job Corps sample, but incomplete information on labor market experience and no information on criminal activities or transfer receipts. Finally, the baseline data would have been defined for 1994 (1 year before the data collected for the treatment sample) for a sample that had already experienced 6 years of attrition, thus threatening the representativeness of the sample.

The other data sets under consideration also had disadvantages sufficient for their disqualification. The SIPP and CPS data sets included the full age span of Job Corps eligibles, but the sample sizes were inadequate (fewer than 1,000 each). Furthermore, these data sets provided no longitudinal data for 36 months or later and no information on criminal activities.[9] The PSID also included a sample of fewer than 1,000 eligible youths and provided only limited information on those who were not heads of households. Finally, the NLSY provided detailed information on a cohort of 4,000–5,000 youths. Unfortunately, these youths were aged 14 to 21 in 1978 and would have been a promising comparison group for Job Corps applicants in 1981: by 1995, however, they were aged 31 to 38.

Other difficulties with using existing surveys for comparison groups are worthy of mention. These difficulties relate to demonstration treatment samples that are not nationally representative or that measure outcomes in idiosyncratic ways (which may reflect limitations of the demonstration rather than the potential comparison sample):

- In recent years, the number of state-based policy evaluations, particularly in the area of welfare reform, has been increasing; similarly, a demonstration of a school reform initiative might be concentrated in one or a small number of states. Existing national databases may lack a large enough sample in the states of interest; furthermore, some data sets may not provide state identifiers in their public use data files.[10]
- Many demonstrations take place in a judgmental (that is, not randomly selected) sample of sites that may not be representative of the national target population for a policy initiative. These demonstrations must confront a methodological tension between identifying a comparison group that is as similar as possible to the treatment population (to promote internal validity of the results) and extrapolating findings to a national target population. To the extent that both variants of a comparison population can be identified in a national data set and their differences measured, use of a national database as a comparison reference, rather than an independent but nonrepresentative

comparison group, could enhance our ability to draw policy implications from demonstrations not conducted with a representative sample.

- Designers of demonstration evaluations often complain that national data sets do not measure potential outcome variables in a manner appropriate for assessing policy impacts of interest. This criticism cuts both ways, however. To the extent possible, program evaluators should attempt to express their findings in terms of broadly available outcome measures in order to promote the interpretability of their results. On the other hand, although certain types of information involving such subjects as criminal activity, drug use, or sexual activity may be inappropriate for broad-based longitudinal data sets, designers of survey instruments for future longitudinal data sets should attempt to incorporate the information required to construct variables that are widely usable as outcome measures for policy evaluations.

The discussion here about the deficiencies of specific data sets relative to an independently tailored comparison group may be moot in an evaluation that rejects a "well-constructed" comparison group in favor of a randomized control group. Most of these issues will remain relevant, however, when discussing the possibility of defining future longitudinal data sets that incorporate or can accommodate a formal experimental design.

Deficiencies of Comparison Groups Relative to Randomly Selected Control Groups[11]

The classical statistical methodology underlying randomized experiments requires that we compare two independent random samples—one that receives the intervention of interest—drawn from the same population. When this condition is met, simple statistical tests reveal the likelihood that any observed differences could be due to chance rather than to systematic differences created by the intervention.

Random assignment fulfills this condition proactively, if neither the sample selection and randomization process nor the method of introducing the intervention creates contaminating effects that could be confused with the intervention's impact. Comparison group methods, on the other hand, use assumptions, measurement of other sources of differences, and statistical models to eliminate differences that could derive from reasons other than the intervention. If these efforts are successful, a residual difference can be identified as resulting from the intervention, perhaps with some measure of statistical confidence.

Continuing debate about whether nonexperimental comparison groups can be used to provide convincing measures of program impacts has been fueled by a number of studies comparing impacts estimates based on control and comparison groups.[12] The debate has also been advanced by an increasingly rich econometric literature about methods to deal with the problem of "selection bias," which results from sources of unmeasured or unmeasurable differences between treatment and comparison groups.[13]

Successful use of nonrandomized comparison groups requires that we be able to measure and control for all systematic differences (other than the intervention) between the samples. Even if all differences can be measured and controlled for, we must keep in mind that the correction process "uses up" statistical power that is no longer available for testing the intervention's primary impact. Time and time again, statistical tests appropriate for randomized experiments are misapplied to nonrandomized comparison groups, with a resulting vast overstatement of the strength of the results.

Similar problems exist with the statistical methods available to test for the presence of and correct for selection bias. Tests for selection bias produce three possible outcomes: 1) bias is present, but we lack an acceptable method to correct for it or perhaps even to detect it; 2) bias is present, and available methods permit us to correct for it; and 3) no systematic bias appears to exist. Each of these outcomes poses problems:

- In the first case, internally valid estimates of impacts cannot be obtained, and the researcher must seek alternative data sets. This is a useful result for researchers evaluating alternative secondary data sets, but scarce comfort for those who have just completed a demonstration with a primary data collection effort.
- In the second case, increasingly sophisticated statistical methods have been developed to correct for the source of bias. However, they typically require the availability of measures for both the treatment and comparison groups that are correlated with program participation but not with program impacts, and tend to produce unstable, nondefinitive results. Even when successful, they absorb statistical power in the correction process and often produce standard errors of impact estimates that are approximately *three times* those produced with demonstrations using control groups. When this happens, sample sizes for a comparison group design have to be as much as *nine times* larger to measure program impacts with the same statistical precision as with a properly designed randomized experiment.
- Only in the last case can we proceed with no statistical correction for bias. Again, however, using the full sample as if random assignment had occurred implies not only that "we have failed to detect evidence of selection bias," but also that "we know with certainty that it is absent."

In any event, we would not know which case applies until a demonstration has been completed and the data have been collected.

The current array of methods available to measure program impacts with nonexperimental data are extremely valuable when time, resources, or other circumstances prevent the design and execution of a randomized experimental design for testing a new policy intervention or an existing program. They are also important for helping to counteract the inevitable imperfections in formal experiments implemented in actual demonstration or program environments.[14] Yet, nonrandom comparison groups—whether "made to order" or drawn from currently available or future longitudinal data sets—are unlikely to return as the methodology of choice for major impact evaluations that place priority on obtaining convincing results. Thus, future longitudinal data sets are unlikely to play a prominent role in impact evaluations unless they can be adapted to accommodate a formal experimental structure for program evaluation purposes. The next section looks at this topic.

Adapting National Data Sets to Accommodate Formal Policy Experiments

If national data sets can be adapted to accommodate formal policy experiments, they could contribute a vital element commonly absent from such experiments: a nationally representative context in which to test a policy.

Internal validity and external validity are two concepts central to sound evaluation design. Internal validity addresses whether what we observe is in fact caused by an intervention. External validity involves whether observed demonstration impacts would be replicated if implemented in broader settings and/or on a larger scale. Although both concepts are crucial for policymakers, it is in the realm of internal validity where well-designed randomized experiments have established their clear superiority over comparison group methodologies. Experiments as typically implemented fall short of standards of external validity, leaving the analyst to engage in nonexperimental, often judgmental methods to establish policy relevance.

An implicit but major controversy in the evaluation community exists between those who focus on establishing an internally valid experimental setting—often by creating an artificial program in an analytically precise environment in one or a small number of nonrepresentative sites—and those who are willing to sacrifice "design rigor" for evaluating a program in a more representative setting. Frequently, researchers face the tension between asking the right question with a weak methodology and asking the wrong question with a sound methodology.

Only recently have there been any significant attempts to place randomized designs in a nationally representative operational setting. The Upward Bound and Job Corps evaluations are prominent examples of these efforts. By providing a national context—or a well-defined target group, such as inner-city students or rural schools—future national databases may provide a vehicle for implementing policy experiments with a presumptive claim of external as well as internal validity for evaluation results.

In the introductory section, I suggested ways in which experimental designs might be integrated into ongoing longitudinal databases: 1) implementing a specific experiment with the initiation of a longitudinal survey; 2) designing a longitudinal survey to accommodate one or more as-yet-unspecified future experiments; 3) augmenting a survey with supplemental sampling units that will receive an experimental intervention, or expanding a longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and 4) providing a sample frame for the random selection of schools for testing school-based innovations. This section provides examples to illustrate the potential and the drawbacks of each of these approaches.

Implementing a Specific Experiment With the Initiation of a Longitudinal Survey

Suppose we wish to test a new approach to enhance reading skills, beginning in the 8th grade for students in inner city schools, and that we are prepared to implement a test of this method that coincides with the initiation of a new NELS-type survey—that is, a longitudinal survey of a random sample of 8th-grade students drawn from a first-stage representative sample of perhaps 1,000 schools. Combining these two initiatives could provide four distinct advantages to the evaluation:

1) With a random sample of students from inner city schools from the NELS frame, the evaluation results could be interpreted directly in terms of the target population (external validity);

2) With common data collected from both students in the demonstration schools and the full sample, the performance of targeted students could be compared with that of their designated control group and that of all students nationwide;

3) With continued tracking of the sample on a longitudinal basis, long-term impacts of the demonstration could be measured beyond the initial evaluation effort; and

4) The incremental cost of the demonstration is likely to be lower than that of a stand-alone study.

The experiment could take one of two general forms—*within-school* versus *across-school* random assignment—each with distinct methodological and operational implications. A demonstration with within-school random assignment of students to treatment and control groups (or more broadly, within-site randomized demonstrations) is the most common design for a policy experiment. A less frequently observed design—but very promising in many contexts in my judgment—involves the random selection of treatment and control *schools*, with all eligible students in the respective groups of schools constituting the treatment and control samples of students.

Demonstrations using within-school random assignment require that the scale of the program intervention in each site be smaller than the potentially eligible population. They also require that the nature of the intervention be such that none of its benefits "spills over" onto the control group, such as when instructional methods for the control group are affected by what teachers learn from the demonstration, or when innovations or reforms are schoolwide in their potential impact.

Within-school designs also require overcoming school resistance to denying program services to some eligible students on a random basis. This resistance increases if there is a risk that some program slots may remain vacant because some applicants are diverted to a control group. The Upward Bound demonstration dealt with this problem by assigning some of the control group (on a random basis) to a waiting list, from which students could be selected to fill vacant slots.

An advantage of using across-school random assignment is that treatment-group schools would not have to deal with the mechanics of random assignment. Control schools would be treated like all others in the longitudinal sample, except to the extent that specialized data collection or an increased sample of students is required.[15]

Innovations tested with this approach must be applied either to the *entire* eligible student population, however, or to subsets of the population identified by student characteristics that can be readily measured in data collected for students in the control schools. Interventions targeted at a small number of volunteer applicants from a larger, nominally eligible group—such as Upward Bound—would not be well suited for this approach, because attempts to identify the comparable group of students in the control schools would suffer from the same selection bias problems that plague nonrandomized comparison groups.[16] Furthermore, the across-school approach could not be used for evaluating existing programs, which are likely to be present already in the control schools.

Finally, in situations for which either design would be methodologically appropriate, across-school designs would typically require larger sample sizes than within-school designs for equal statistical precision, because both individual and school characteristics would vary randomly between the treatment and control groups. Within-school random assignment, on the other hand, eliminates variations in school characteristics between the treatment and control group.[17]

The issues discussed here involving the choice between within-school and across-school randomized designs are relevant whether or not a demonstration is integrated into a longitudinal survey. Special problems to be considered when integrating either approach into a longitudinal survey include the following:

- We must have identified the experiment of interest in time for implementation at the beginning of the longitudinal survey. *More importantly, the target age cohort for the experiment must coincide with a cohort included in the survey.* If the survey is tracking a cohort of 8th graders, for example, a demonstration targeting that group could be included, but not one focusing on 10th graders.
- Planners of demonstrations typically solicit applications from schools or sites willing to participate. The strategy discussed here requires approaching a random sample of schools and inviting them to participate. This approach is feasible only if the offer of participation is sufficiently attractive to achieve high participation rates.
- The number of students per school in an NELS-type survey is unlikely to be large enough to support the sample-size requirements of a demonstration. Thus, the sample of students would have to be augmented in the treatment schools and probably in the control schools (in the across-school design) as well.
- The content of the longitudinal survey may have to be modified to ensure that it includes appropriate outcome measures for evaluating the long-term impact of the intervention. In addition, supplemental data collection may be required for the demonstration (for example, if achievement test scores are desired).
- Students in the treatment sample, *by virtue of their receipt of the program innovation*, would no longer be representative of their cohort. Thus, the size of the longitudinal sample, excluding the demonstration sample, would have to be large enough to serve the general purposes of the longitudinal survey.

Designing a Longitudinal Survey to Accommodate Future Experiments

Simultaneous initiation of a randomized demonstration and a longitudinal survey requires that an uncomfortably large number of planets be in proper alignment. The increased flexibility of a longitudinal survey that could accommodate one or more experiments *after* its initiation would be desirable. For example, we may want to test an initiative targeted at 10th-grade students

2 years after the beginning of the longitudinal survey, or we may not yet have settled on a policy initiative worthy of experimentation.

What characteristics must the survey sample have to offer this flexibility, and what special problems would have to be resolved in designing subsequent evaluations? For the purposes of this discussion, assume that the longitudinal survey would track a cohort of 8th-grade students, with follow-up interviews scheduled every 5 years.

Several general issues would have to receive special attention in the design of the longitudinal sample to accommodate future experiments, some of them already identified. First, the questionnaire content might have to be examined in terms of its measurement of student characteristics and outcome variables likely to be important for evaluating future policy demonstrations. If the demonstrations require supplemental data collection, especially on a continuing longitudinal basis, much of the advantage of attempting to integrate the demonstration with the survey would be vitiated.

Second, we would have to consider the available sample sizes for all potential evaluation target groups, both for potential demonstrations and for the remaining sample available for general users of the longitudinal database. Realistically, most strategies for appending a demonstration would involve adding supplemental samples (both school sampling units and students within schools) to the survey at the time the demonstration is implemented.

Third, a survey like NELS, restricted to a single-grade cohort, would be particularly restrictive in terms of the future timing and range of potential demonstrations. For example, a longitudinal sample of 8th graders could be integrated with a policy initiative directed at high school sophomores after 2 years, but not at any other time. A survey with more than one cohort would be more flexible in terms of its potential accommodation of future demonstrations.

Finally, if inclusion in the longitudinal sample places schools or students "at risk" of inclusion in a future demonstration, there may be issues of informed consent to consider. (This is more likely to be a problem for demonstrations calling for within-school random assignment than for the across-school approach, if responding to an interview increases an individual's exposure to future selection for participation in an experiment.) Such consent, if required, could lower response rates in the longitudinal survey, a problem that compounds in subsequent waves of the survey. Again, this problem is mitigated if we think in terms of supplemental samples for demonstration implementation.

Returning to the example of testing a policy initiative targeted at high school sophomores 2 years after a longitudinal survey of 8th graders has been initiated, the designer of the demonstration would face several obstacles:

- Timing is everything, as already suggested. Two years after the initiation of a longitudinal survey focused exclusively on an 8th-grade cohort is the *only* time a demonstration targeted at 10th graders could be implemented.

- In this example, baseline data are 2 years old and would not exist for any augmented sample required for the demonstration. If baseline data in addition to student records are required, a supplemental baseline survey would have to be implemented.[18]

- During a 2–year period, students in the longitudinal sample may have dispersed to different high schools in their districts, moved out of the area, dropped out of school, or otherwise disappeared from the sample. Sample students remaining in the same school districts would not be representative of all students in those districts, because students who changed districts in the past 2 years would be excluded from the sample. These factors would severely complicate attempts to implement a demonstration using students already included in the sample, even if sample sizes available for the demonstration were adequate.

- If a supplemental sample is drawn for a demonstration treatment group, the above complications could compromise the suitability of using the regular longitudinal sample in a selected set of schools as a control group. A potential solution to this problem might include adding a supplemental sample of control students who arrived in the sampled schools since the definition of the longitudinal frame.

- Demonstrations that combine the "new" and "original" sample would have to deal with potential differential sample attrition over time, resulting from the different "longitudinal ages" of the two portions of the sample.

These considerations are likely to make separately drawn treatment and control samples more attractive to program evaluators than designs relying heavily on the "original" sample from a previously initiated longitudinal sample. The question is whether these samples should retain a structural link to the longitudinal survey, or whether the current practice of implementing randomized demonstrations independently of national longitudinal samples should continue.

Augmenting a Survey With Supplemental Sampling Units to Receive an Experimental Intervention, or Expanding a Longitudinal Sample to Incorporate Separately Defined Demonstration Treatment and Control Groups for Common Data Collection Efforts

The discussion here has implicitly moved us in the direction of a more limited integration of randomized demonstrations with longitudinal data sets. Three possibilities come to mind:

1) Augmenting the longitudinal sample with supplemental sampling units—*selected in the same manner as schools forming the basis of the longitudinal survey*—to receive the program intervention being evaluated, and using the longitudinal sample as a control group;

2) Choosing supplemental sampling units (in the same manner) for *both* the treatment and control groups, but integrating the demonstration into the longitudinal data collection sample; and

3) Defining a demonstration sample by procedures not related to the longitudinal frame, as is currently done, and limiting the link to common longitudinal data collection.

The first approach forces the treatment sample to be nationally representative of the target group in question, a major advantage over most contemporaneous randomized demonstrations. This approach would also be viable as a variant of the first scenario described in this section, in which a demonstration is implemented at the same time the longitudinal survey is initiated. When the demonstration is initiated *after* the longitudinal baseline, however, a number of issues related to the comparability of the treatment and control groups (discussed earlier) could compromise the experiment's validity.

The second approach would utilize supplemental, representative sampling units for both the treatment and control groups. Although not using the longitudinal sample in a literal sense, this approach combines the advantages of providing a nationally representative test of the program intervention on a sample defined in the same manner as that used to track students nationally, with the economic advantages and the interpretative consistency of commonly collected data for demonstration participants and the general student population. This approach could strengthen program evaluation methodology significantly and is an option worth pursuing where feasible.

The third option leaves demonstration designers free to define independent treatment and control groups, while retaining the advantages of common data collection efforts. This approach may be an improvement over current practice, but I would find it to have a rather disappointing outcome: guiding randomized demonstrations in the direction of nationally representative rather than pragmatic implementation venues, which would be achieved by the previous option, is an important priority for the evolution of program evaluations.

Providing a Sample Frame for the Random Selection of Schools for Testing School-Based Innovations

All the design options discussed here have been "school-based" but focused on measuring outcomes through longitudinal data collection efforts patterned after NELS, with schools serving as the primary sampling unit for selecting students and as the venue for implementing the demonstrations. The tested policies were viewed as affecting specific students enrolled in the demonstrations, rather than as broader school reforms that might have schoolwide impacts.

Here, the discussion expands to include experiments in which the school is the target of the innovation, and the design is clearly across-school in character. Measured outcomes might take the form of longitudinal observations of students, as before, or repeated outcome measures for successive cohorts of students in a longitudinal sample of schools. In the latter case, measured outcomes could be based on administrative records, test scores, or aggregate measures for each school, as well as student interviews.

The design objective here is to use existing survey sample frames to select random samples of schools for testing a reform or innovation in a formal experiment, rather than to follow the more traditional approach of comparing judgmental treatment and comparison samples of schools.

For example, suppose we wish to test the effect on mathematics achievement or other outcomes of making personal computers readily available to students in rural schools.[19] In order to implement such a demonstration, we would select a random sample of rural schools from the NCES Schools and Staffing Survey (SASS) frame (or augment the sample if there are not enough schools) and invite these schools to participate in the demonstration.[20] Rural schools not selected for the offer of participation would constitute the control group and could be augmented by an additional sample of schools, if necessary. Outcomes could be measured with supplemental data collection efforts in conjunction with future waves of SASS.

In order for experiments of this sort to be effective, certain conditions would have to be met:

- The SASS design would have to be modified to be more longitudinal in character. (It is my understanding that such a modification is under consideration.) Furthermore, it would be desirable to investigate the possibility of adding summary outcome measures relevant for a range of potential school innovations and reforms to limit the extent of supplemental data collection efforts.
- As noted, the tested initiatives would have to be attractive enough that a large proportion of the selected schools would agree to participate, because the treatment group would be properly defined as all who are offered participation (not just participants). Furthermore, nonparticipants would dilute the power of the experiment.
- The potential impacts of the intervention would have to be schoolwide or serve a high fraction of identifiably eligible students. The impacts would also have to be measurable in tangible terms that could be measured consistently across schools.
- If all relevant output measures could be obtained from standard survey data, there would be no need to obtain any special consent from the control schools. Agreement of control schools to participate in supplemental data collection efforts would have to be solicited, but they would not have to be involved in the demonstration in any other material way.
- If the initiative is widely publicized, inexpensive, and easy to implement, there is the risk that control schools will implement a similar program on their own. If this happens too quickly, the outside world will "catch up" to the innovation before its impacts can be measured. The demonstration is more likely to be successful in measuring impacts if the innovation requires significant resources and/or technical assistance to implement, and if premature publicity surrounding the demonstration is kept to a minimum.

SASS may be less promising (in terms of its structure and traditional content) than longitudinal student samples as a vehicle for collecting required outcome measures, increasing the likelihood that specialized data collection efforts would have to be implemented in conjunction with randomized demonstrations. Even if supplemental data collection is required, however, I place high priority on the possibility of executing randomized tests of school-based interventions within the standardized framework that a nationally representative database can provide.

CONCLUSIONS: A REALITY ASSESSMENT

In the discussion here, I reviewed a range of reasons why demonstrations and policy evaluations have not made significant use of existing national data sets, and considered a number of ways in which these data sets might be adapted to alter the conduct of future evaluations. My conclusions are as follows:

- The descriptive value of well-structured national data sets for framing policy issues ought not to be minimized, and precautions should be taken so that efforts to accommodate policy experiments do not dilute this value.
- Attempting to improve the attractiveness of national data sets as general-purpose (nonrandomized) comparison groups would not be, in my opinion, a noble objective. The intrinsic weakness of comparison groups relative to randomized control groups makes this effort an unpromising investment.
- Experiments with within-site treatment and control groups are difficult to incorporate into a national longitudinal sample, unless timing of a demonstration implementation converges fortuitously with initiating a longitudinal survey that has a compatible age cohort. There is potential for improving the efficiency and comparability of longitudinal data collection, however, and for moving such experiments in the direction of using representative sample frames compatible with national data frames.
- Attempts to append an experiment to a longitudinal survey after the survey's initiation point would be fraught with difficulties, unless supplemental samples are drawn for the demonstration treatment and control groups.
- The potential for implementing demonstrations with across-school random assignment appears to have been severely underestimated, both for student- and school-targeted initiatives. Coordinating the design of such evaluations with the representative frames of national surveys and engaging in integrated data collection activities, where possible, could produce significant improvements in both the methodology and the efficiency of future policy experiments.

NOTES

1. The evaluation strategies for both Upward Bound and Job Corps depended on pools of potential eligibles that exceeded the available number of program slots.

2. SIPP utilizes an overlapping panel design rather than a strict longitudinal design; the Current Population Survey, another widely used continuing data set, utilizes overlapping panels of household locations. The NCES Schools and Staffing Survey, which will be considered later in this paper along with the NCES longitudinal studies for adaptation to experimental evaluations, utilizes repeated cross-sections with an approximate 30 percent overlap of schools between successive interview waves.

3. This paper defines a "control group" as a sample selected through random assignment between treatment and control students or schools, and a "comparison group" as one chosen to be as similar as possible to a treatment group, but without random assignment.

4. An additional potential focus for incorporating experiments into NCES national data collection efforts—not the subject of this paper—would involve testing alternative data collection methodologies. Designing these experiments would involve substantive issues related to the data collection methodologies being tested, but the sampling and experimental design issues would be relatively straightforward.

5. For example, the initial design for the ED's evaluation (conducted by Mathematica Policy Research) of the tech-prep educational program called for using data from the National Education Longitudinal Study as a comparison group, but this approach was abandoned after critical examination by both Office of Management and Budget and project staff. NELS was one of several longitudinal data sets considered for creating a comparison group for DOL's Job Corps evaluation before a randomized design was chosen as a superior approach.

6. School-based interventions would require longitudinal samples of *schools*, but associated samples of students might appropriately be repeated cross-sections, depending on the evaluation.

7. The upper age limit was increased from 21 to 24 in 1993.

8. A previous evaluation of Job Corps completed in 1982 (Thornton et al. 1982) identified a reduction in criminal activities as a prominent benefit of the program.

9. The rotation pattern of the SIPP panels provided longitudinal data for 30 months or less; the CPS utilizes rotating panels based on household location and provides no actual longitudinal data.

10. NCES usually maintains both public use and restricted use data files. The restricted use files may permit identification of states and other locations.

11. Portions of the following discussion are adapted from Metcalf and Thornton (1991).

12. See Ashenfelter and Card (1985), Lalonde (1986), Lalonde and Maynard (1987), and Fraker and Maynard (1987).

13. For example, see Maddala and Lee (1976), Heckman (1979), and Heckman and Hotz (1989). For discussions of effective use of tests and corrections for selection bias in nonexperimental data, see Heckman and Robb (1985) and Heckman et al. (1987).

14. For example, the potential presence of selection bias must be dealt with when 1) fewer than 100 percent of the individuals selected for a treatment group choose to participate in a program; 2) separate impact estimates are desired for different program elements provided to nonrandom subsets of the treatment population; and 3) longitudinal data collection efforts produce differential attrition rates for the treatment and control groups.

15. In the limiting case, all schools in the longitudinal sample with the characteristics of the treatment schools—in this example all inner-city schools—would be part of the control group by virtue of their inclusion in the longitudinal sample. The control group schools have no special knowledge about the existence of the demonstration in the treatment schools.

16. Interventions that require voluntary enrollment could be tested if the participation rate is high (for example, 70 percent or greater). The defined treatment and control groups, however, would include all eligibles, inclusive of nonparticipants. The presence of nonparticipants would dilute the precision of measured impacts by a factor proportional to $(1/P^2)$, where P is the participation rate.

17. By eliminating this major component of variance, within-school random assignment improves the statistical precision of internally valid estimated impacts for the demonstration schools. Extrapolations to national estimates would still have to account for design effects due to the clustering of the student sample into a small number of schools, but a within-school design would retain its statistical advantage for extrapolations as well.

18. In principle, baseline data are not required for comparing treatment and control outcome data in a properly constructed experiment. Baseline data, however, can be used to reduce the variance of impact measures by controlling for student characteristics, and can be invaluable for interpreting future problems of sample attrition. Student records might serve some of this purpose, if informed consent issues can be resolved.

19. Alan Hershey of Mathematica Policy Research suggested this example. Recently it has come to my attention that a similar program already exists.

20. Alternatively, one might choose schools serving as primary sampling units in a longitudinal student survey, but this approach would provide relevant student data only if the demonstration were implemented at the initiation of the longitudinal survey, and only if the survey included a broad enough age span of students to encompass the intended target of the reform.

REFERENCES

Ashenfelter, O. and Card, D. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67 (4).

Fraker, T. and Maynard, R. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* 22 (2).

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47.

Heckman, J. J. and Robb, R. 1985. "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data.* Eds. James J. Heckman and Burton Singer. Cambridge, MA: Cambridge University Press. 1985.

Heckman, J. J., Hotz, V. J., and Dabos, M. 1987. "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" *Evaluation Review* 11.

Heckman, J. J. and Hotz, V. J. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84 (408).

Lalonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76.

Lalonde, R. J., and Maynard, R. 1987. "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Review* 11 (4).

Maddala, G. S. and Lee, L. 1976. "Recursive Models with Qualitative Endogenous Variables." *Annals of Economic and Social Measurement* 5.

Metcalf, C. E. and Thornton, C. 1992. "Random Assignment." *Children and Youth Services Review* 14(½): 145–156.

Thornton, C., Long, D., and Mallar, C. October 1982. "A Comparative Evaluation of the Benefits and Costs of Job Corps After Forty-Eight Months of Post Program Observation." Princeton, NJ: Mathematica Policy Research, Inc.

U.S. Department of Education, Office of Educational Research and Improvement. January 1995. *Programs and Plans of the National Center for Education Statistics, 1995 Edition.* Washington, D.C.: U.S. Department of Education.

Discussant Comments

DONALD B. RUBIN

I congratulate Chuck Metcalf for writing a clear and direct article advocating the increased use of randomized experiments in educational research, a point with which I fully agree. He is also to be congratulated for providing a list of good recommendations on how to conduct such studies (e.g., by imbedding them in longitudinal studies and doing treatment assignment at an appropriate level to avoid issues of interfering units). It is especially rewarding to see a distinguished, practically experienced economist strongly eschew the naive application of simple OLS models, structural equations methods, and instrumental variables techniques that have been advocated by many in economics (e.g., Heckman 1979, and other more recent references cited by Metcalf).

I am particularly interested in his citation of LaLonde (1986) to support his advocacy of randomized experiments because that article has become a focal point in a course on "Causal Inference," which I have been teaching with Professor Guido Imbens in the Department of Economics at Harvard University. Specifically, the LaLonde article shows that the standard techniques typically used by statisticians and economists with nonrandomized data cannot be trusted to provide the "correct" answer, where correct is defined by the answer provided in a randomized experiment. In this study, the treated group, consisting of about 200 from a randomized experiment concerning a job training program, was considered as the treated group in an observational study, whereas the comparison group was to be derived, as typical in such observational studies, from a large-scale database (e.g., either the CPS or the PSID). Estimates of the treatment effect were then obtained using the standard array of econometric/statistical modeling tools on the actual treated units and the observational comparison units. These tools provided answers that were typically wild, and often absurd, when compared to the benchwork estimate available from the randomized experiment. The conclusion, which is I believe consistent with Metcalf's position, is that this documents the fact that such observational studies cannot be trusted to produce honest policy-relevant estimates of treatments.

When Imbens and I presented this example in class, it was in the context of already having warned the students of the extreme extrapolation often implicit in estimates based on such methods, and of already having exposed them to propensity score methods (Rosenbaum and Rubin 1983, 1984, and 1985; Rubin and Thomas 1992a, 1992b, 1996), which avoid such extrapolation. Propensity score methods can also directly lead to the conclusion that, despite the apparent wealth of comparison information available in large databases such as the CPS and PSID, the treated and comparison groups may be so far apart that there are no trustworthy conclusions possible. Two economics students, Sadek Wahba and Rajeev Dehejia, pointed out that the conclusion from LaLonde, to the effect that such studies are hopelessly unreliable, should be decomposed into two crisper issues. First, are the *data* from studies such as LaLonde's hopelessly unreliable? Second, are the *standard methods* used to analyze such data hopelessly unreliable? We all seem to agree that the latter is true, but what would happen if LaLonde's data were reanalyzed using the far more appropriate propensity score methodology, now very popular in much of social science and medical research (e.g., U.S. GAO Report, "Breast Conservation Versus Mastectomy," 1994).

LaLonde graciously supplied the data, and Sadek and Rajeev went to work. Despite the thousands of potential control units in these large-scale data sets, only about 200 or fewer were similar enough to the treated groups, with respect to their propensity scores, to be considered to constitute a reasonable comparison group for the treated. Adjustment then took place using special versions of either subclassification (Rosenbaum and Rubin 1983) or matching (Rosenbaum and Rubin 1985) on the propensity scores, with possibly some simple OLS regression for minor adjustments. Of great importance, the results based on the propensity score technology tracked those from the randomized experiment, even with respect to interactions between treatment and some background characteristics. An initial reference for this project is Wahba and Dehejia (1996).

Certainly this work does not show that using propensity score techniques in observational studies *will* always either 1) conclude the treated and comparison groups are too far apart, or 2) provide an estimate like that from a randomized experiment. But Wahba and Dehejia (1996) provide an important "existence theorem," showing that propensity score technology, because it inherently addresses problems of extrapolation, *can* produce acceptably accurate estimates of causal effects from observational data in cases where the standard OLS or selection model methods fail to do so.

My conclusions, therefore, are a tempered version of Metcalf's. That is, we should push for randomized experiments whenever possible, but because observational data will nearly always be cheaper to obtain and more readily available, we should be willing to analyze nonrandomized data, but with great care, using appropriate propensity score methods and avoiding unreliable model-based extrapolations employing standard statistical or econometric models. These models have their place, and the ideas underlying some of them can be extremely useful in some contexts (e.g., see Angrist, Imbens, and Rubin 1996); however, they must be used insightfully and not be used, as often advocated, as off-the-shelf solutions to the problems of possible "selection bias" in observational studies.

References

Angrist, J.D., Imbens, G.W. and Rubin, D.B. Forthcoming 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, Applications Invited Discussion Article.

Heckman, J.J. and Robb, R. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*. Eds. James J. Heckman and Burton Singer. Cambridge, UK: Cambridge University Press.

LaLonde, R.J. 1986. "Evaluating the Econometric Evaluations of Training Programs With Experimental Data." *American Economic Review* 26.

Rosenbaum, P. and Rubin, D.B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Rosenbaum, P. and Rubin, D.B. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79: 516–524.

Rosenbaum, P. and Rubin, D.B. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Incorporating the Propensity Score." *The American Statistician* 39: 33–38.

Rubin, D.B. and Thomas, N. 1992a. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *The Annals of Statistics* 20 (2): 1079–93.

Rubin, D.B. and Thomas, N. 1992b. "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Covariates." *Biometrika* 79 (4): 797–809.

Rubin, D.B. and Thomas, N. Forthcoming 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." Forthcoming *Biometrics*.

U.S. GAO Report to the Chairman, Subcommittee on Human Resources and Intergovernmental Relations, Committee on Government Operations, House of Representatives. 1994. "Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies."

Wahba, S. and Dehejia, R.H. 1996. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Cambridge: Department of Economics, Harvard University.

6 Postsecondary Education

Tracking the Costs and Benefits of Postsecondary Education: Implications for National Surveys

---

Michael S. McPherson
Morton O. Schapiro

INTRODUCTION

Our assignment is to advise NCES on ways their data collection activities could help shed more light on understanding the costs and benefits of higher education. This paper begins with a discussion of how educational impacts are identified and valued and then goes on to distinguish between the immediate and long-run consequences of educational investments. This discussion is followed by an explanation of what we mean by "high quality" data in arguing for the importance of certain types of longitudinal data sets. The next section addresses the role of "educational treatments" in identifying how educational efforts and resources translate into impacts on students' learning and concludes with a discussion of the usefulness of simple cost/benefit measures in international comparisons of educational "productivity."

CAUSAL VERSUS EVALUATIVE ISSUES

Appraising the costs and benefits of postsecondary education requires knowledge of the impacts of such education—a problem of identifying causation—and knowledge of the values to be placed on those impacts.

Difficulties in Identifying Causal Impacts

It is often relatively easy to identify *differences* between people who have and who have not attended college, or even among those who have had different types of postsecondary experience. But it is much harder to identify the causes of these observed differences among, for example, college graduates and high school graduates.

Two major statistical problems that make causal analysis in this area difficult are *maturation effects* and *selection effects*. Maturation effects create an important hazard for individuals who try to reflect on how their college experience affects their own lives. Looking back, it may be easy to identify ways in which one was different after college than before attending college. But to some unknown extent, those differences, rather than being caused by the college experience, may have been simply a result of aging. In studying individuals, it is hard to surmount this problem of distinguishing the effect of the college experience from the simple effect of the passage of time. This is one basic reason for attempting to assess the effects of college by comparing college-goers to non-college-goers, rather than simply looking at changes occurring in the lives of people who did attend college.

But comparing college-goers to non-college-goers raises the problem of selection effects. These arise because the processes that determine who goes to college, as well as who goes to which college, are far from random. A great deal of evidence shows that college-goers differ systematically at time of entrance from non-college-goers. College-goers come from families with higher incomes; they score higher on average on aptitude tests; they are more likely to have parents who attended college, and so on. An important advantage of rich longitudinal databases tracking individual life histories—such as the National Longitudinal Study of 1972 (NLS–72), High School and Beyond (HS&B), and the National Education Longitudinal Study of 1988 (NELS:88)—is that they allow us to observe and statistically control for many of these differences in estimating statistically the impact of college experiences on later life.

Unfortunately, however, no data set is rich enough to enable us to observe all the ways in which college-goers differ from non-college-goers (or the ways in which people with different postsecondary education experiences differ from one another). To the extent that these unobservable differences between college-goers and non-college-goers themselves lead to differences in what we observe about people in their later lives, we are at risk of mistakenly attributing these later differences to the college experience, rather than to the unobservable differences that led one group to attend college while another did not. Econometricians have spent considerable energy and imagination in finding ways to allow statistically for these selection effects, and much progress has been made. Still, selection effects remain a great obstacle to sorting out the causal impacts of college-going.

Even more difficult than measuring the effects of college is understanding why or how those effects occur. Better knowledge of the effects of college on people's later lives might help guide decisions by individuals about whether to attend college or by governments about whether to encourage college attendance. But knowing what *features* of a college experience lead to particular outcomes would be of great help in permitting colleges to improve their operations. Clearly to study such problems requires a much more fine-grained measurement of various dimensions of the college experience than most existing data sets permit.

Difficulties in Evaluating Outcomes

Cost-benefit analysis requires not only the identification of outcomes but also the evaluation of those outcomes, and of the inputs that produce them in systematic and preferably quantitative ways.

Benefits of postsecondary education appear partly in labor markets. It is commonly believed that postsecondary education equips people with skills and knowledge that make them more productive. Such higher productivity may then turn up in higher wages, so that the wage differences between, say, college graduates and high school graduates are an index of the social benefits of higher education. But even when the focus is limited to labor market effects, this analysis is not so simple. First, of course, selection effects like those just noted imply that differences in wages between high school and college graduates are probably not all due to the effects of college. Indeed, it is possible to develop a coherent economic model of college in which the economic function of higher education is to sort out more and less productive individuals, rather than adding to their individual productivity. But even when wage differences result from changes in individual productivity caused by college, these wage differences may either understate or overstate the impact of college on economic productivity, simply because wages may understate (e.g., school teachers, public defenders) or overstate (e.g., investment bankers, deans) the social contributions of particular jobs.

Quite apart from labor market effects, higher education may make people more valuable in other ways, as by making them more politically active or more community minded. These effects are hard to measure in a causal sense, and even harder to measure in a cost-benefit sense, since putting dollar valuations on such effects is difficult.

It is also important to note in passing that higher education makes major contributions to social productivity through its contributions to knowledge accumulation and basic research. Examining ways to improve measurement of these benefits and of the cost of producing them is beyond our scope here.

Unlike the measurement of benefits, the measurement of costs may appear straightforward. But it is actually more difficult than it may appear. One difficulty is that of attributing costs to particular activities and hence to particular outcomes. It is, for example, quite difficult to separate the costs of graduate and undergraduate education in most existing data sets. (There are, of course, conceptual problems in trying to allocate those university costs that contribute jointly to undergraduate and graduate education, but even simple measures, like the number of graduate and undergraduate courses taught by faculty members, are quite hard to come by.)

Another difficulty is sorting out private and social costs and being clear about who pays. The price charged at most colleges, and especially at public colleges, is well below the cost of production. At private colleges, gifts from alumni (often accumulated in endowments) and at public colleges, appropriations from state governments, keep the price to families down. Thus, the calculus of whether college pays in a cost-benefit sense for the family is quite different from the question of whether it pays for society.

It is also important in measuring private costs to gain clarity about what the student and family actually pay. Because of the importance of financial aid, both grants and loans, the actual costs of attending a particular school may be quite different for different individuals. Further, the most important cost of college for most people is not the out-of-pocket price paid to the school, but rather the opportunity cost of student time—the earnings forgone by reducing or eliminating work hours to attend school.

What is the role of NCES in contributing to these evaluative questions? We would underline the importance of NCES recognizing the limitations imposed on it by its role as a government statistics-gathering agency. Actually putting dollar values on various benefits (and to a lesser extent on costs) is ultimately a political decision—a public decision about values. The job of NCES is to provide the information to support that public decision. So, for example, it would be a contribution if NCES studies could shed light on the impact of postsecondary education attendance on the likelihood of one's participation in volunteer public service activities; it would not be smart for NCES to attempt to put a dollar valuation on the worth of such service contributions.

THINKING THROUGH IMPACTS OF POSTSECONDARY EDUCATION

In studying the impacts of higher education on individuals, it is important to distinguish relatively direct and immediate educational impacts from the long-run effects on earnings and quality of life that are the ultimate payoffs of higher education.

Immediate Educational Impacts

Typically studies of educational assessment and educational production functions assume that education aims at certain impacts on knowledge and cognitive capacities that are thought to be directly related to educational inputs. NCES has strengths and weaknesses in developing data for these kinds of studies.

The basic strategy of such studies is to relate variation in educational inputs to available outcome measures. Abstractly, the ideal framework for such a study is an experiment: introducing planned variation in an input of interest, while applying different levels of the input randomly to a set of students. The fact is, however, that experimental studies of this kind are relatively rare. Much more common are "natural experiments," where naturally occurring variations in inputs of interest are related to corresponding variations in outcomes.

The principal strength of NCES for such studies lies in its ability to develop reliable comparative data for different institutions. Having comparative data across institutions is valuable because it allows for more variation in both inputs and outputs than one is likely to observe within a single institution or a narrowly confined set of institutions. NCES longitudinal surveys like High School and Beyond have enough reach to incorporate institutions with widely varying input levels—large versus small average class size, rich versus meager library resources, and so on.

The principal weakness of NCES here is the counterpart of its strength: the impracticality of generating in-depth data for individual institutions. Two different students at the same institution may have sharply different educational experiences. Without being able to track such variations internal to institutions, educational production function or outcome studies will inevitably involve *averaging* over both the input levels and the outcomes experienced by different students.

If one considers the existing longitudinal surveys (NLS–72, HS&B, and NELS:88), it is clear that they are richer in their information about individual students than in their descriptions of educational environments and inputs at the postsecondary level. Thus, these surveys report the results of student performance on a battery of tests at high school completion and in later years and provide some data on the quality of performance in college—GPA and the like. Information on the learning environment—class size, pedagogical techniques employed, characteristics of faculty—are not a focus of study, and can be inferred, if at all, mostly through linking the survey data to information from the IPEDS financial statistics survey, which itself tracks only very general institutional characteristics, such as spending, in broad categories. And as noted, these surveys do not permit any tracking of differences among students within a school on the educational inputs directed toward individual students.

These limitations are not surprising, and imply no criticism of the existing surveys. The basic fact is that the longitudinal surveys have not been designed principally with the goal of studying direct educational impacts at the postsecondary level. It is important to appreciate that, as a result, they are poor instruments for this purpose. And because this imposes a real gap in our knowledge of the causal consequences of postsecondary education, it also limits the value of these surveys in studying the costs and benefits of higher education. The discussion below will focus on what kinds of efforts NCES could make to address these limitations.

Long-Run Consequences of Educational Investments

We have just been noting limitations on the ability of existing surveys to shed light on what actually happens to students as a direct consequence of their educational experiences. Fortunately or unfortunately, much economic analysis of the long-run effects of college attempts to measure these effects while sidestepping completely the question of how those effects are produced. This is of course the model of the classic "human capital" study, which attempts to measure the private and social returns to education while treating the educational treatment itself as a "black box."

Many studies of the returns to education have treated the basic unit of education as the "year," and have viewed the returns to an added year of education simply in dollar terms, comparing the earnings of those with more and less schooling. This formulation makes the educational input a homogeneous commodity, the "year," and makes the educational output another homogeneous commodity, "the dollar earned." Much has been learned from models that employ such radical simplifications, but plainly much is also omitted. In particular, such studies are worthless from the standpoint of asking how to improve education—whether one type of education or one way of "doing education" is more valuable than another.

More recent studies have added complexity to this simple model of educational effects. On the input side, there are attempts to recognize that educational inputs differ in their intensity (measured, for example, by dollars spent per student on instruction), as well as their duration (measured by years of school). Researchers have attempted also to measure the returns to different types of schooling—public versus private, community college versus proprietary vocational school. Studies of this kind are potentially of great importance in guiding decisions about public investment, which of course raises the stakes in ensuring that such studies can be conducted reliably. On the output side, there are efforts to recognize that the impact of college experience may show up in places other than the paycheck—in choice of vocation and in the various non-pecuniary dimensions discussed above.

Plainly attempts to move in these directions, recognizing the multidimensionality of both educational inputs and outputs, raise data demands rapidly. On the input side, one runs into the problem discussed in the previous section that existing longitudinal surveys have only quite gross measures of educational characteristics of institutions and provide virtually no information on differences in the educational inputs applied to different students. On the output side, the surveys are richer, since they include significant attention to attitudinal questions and to activities outside the workplace. It is our sense that these dimensions of the data in the existing national surveys may have been underexploited. (We think this is clearly true of work by economists, but are less well informed about work done with these data in other academic specialties.)

A major question here is that of the *pathways* through which college experiences influence activities in later life. Consider, for example, evidence that college graduates are more likely to participate in community service activities. This could just be a selection effect—that people who are more likely to engage in public service are also more likely to attend college. But even if the result is not spurious in this way, interpreting it remains a complex matter. Is this because college has changed their attitudes—increasing the value they attach to public service; changed their *capacities*—so that they are asked to do more because they do it more effectively; or changed their opportunities—so that they are offered more interesting or rewarding service opportunities owing to their higher status as college graduates? It is far from clear that survey data can help much in answering these questions, but they are certainly important ones to keep in mind in evaluating research findings.

THE NEED FOR GOOD LONGITUDINAL DATA

As noted earlier, the ideal way to study both short- and long-term effects of college experiences would be through conducting experiments involving random assignment of subjects. Without discounting the possibility of doing this in some settings, it is clear that most knowledge about college effects will not come from this source. Rather, we are thrown back on the "natural experiments" generated by the educational system.

We can make no more important point than that high-quality longitudinal data is an essential component of reliable studies of college effects in non-experimental settings.

This point applies to studies of immediate effects of college experience on student attributes as well as on long-run studies of educational impacts on life outcomes. For the study of direct effects of college, longitudinal data provide benchmark information on student attributes before or at the time of college entry in order to examine how college *changes* these attributes. It is further necessary to make comparisons between the changes experienced by those with and without college experience to distinguish college effects from maturation effects.

More subtle distortions can also be corrected with adequate longitudinal data. Suppose, for example, that a group of students enter college with scores on cognitive tests equal to those of a group of non-college-bound students, and 4 years later the college students have improved their scores more than the non-college-bound. Since we do have pre-treatment data, can we safely attribute the difference to the college experience? Not necessarily, for the non-college-bound, even with the same test scores, may differ from the college-bound in ways that are not picked up in the test score data. They may, for example, attend college because they are more motivated, or because they have reason to believe they will learn faster, and so on. A rich longitudinal data set that tracks pre-college differences among youth may provide measures of variables that correlate with unobserved differences like those just noted, allowing statistical control for these differences that will otherwise confound results. Ultimately, there is no sure cure for such unmeasured effects except random assignment, but good longitudinal data are helpful.

In studies of long-run effects of college, pre-college data are needed for all the same reasons. Post-college data are quite valuable as well. Obviously, one must have data for that point at which the long-run effects of college are measured. And indeed, one can do good work limiting one's data to such "end-point" information. In such work, one is lumping together all the very different kinds of effects college may have on life outcomes, and all the different pathways through which these effects may operate. This sort of "reduced form" or "black box" approach is legitimate, but limited. With good data, much can be added by a more "sequential" approach. Thus, for example, a particular type of college experience may increase a person's likelihood of attending a professional school after college, thereby influencing future vocational choice and career opportunities and ultimately earnings. Sorting out such causal pathways can be instructive in ways that simple bottom-line assessments of the impact of college are not.

What is "Good" Longitudinal Data?

Several times we have referred to the value of "good" longitudinal data. This section concludes by being more specific about what we mean by good data in this context.

Obvious statistical requirements include sample sizes that are adequate to the levels of detail in the analyses that are contemplated and high, preferably uniformly high, response rates among population subgroups. We would also stress three desiderata that are more specialized to longitudinal surveys.

First, it is very helpful to minimize reliance on recall. For pre-college information, this points to the advantage of beginning surveys when subjects are young, so that contemporaneous information on their background, environment, and characteristics can be collected. For post-college information, this points to the advantage of reasonably frequent resurveys in order to minimize reliance on recall to fill in the gaps. The reason for avoiding reliance on recall is obvious: recall is not just imperfect but frequently biased, as the hundreds of thousands of people who claim to have seen Don Larsen's World Series perfect game in person would, unfortunately, not attest.

Second, it is important where possible to cross-check individually reported data against administrative records. One striking illustration of this point is provided by the Postsecondary Educational Transcript Study, which recovered data from colleges on the educational records of students in the NLS–72 study. The re-study turned up large inaccuracies in student reports of their transcripts. The NPSAS studies of student aid do an excellent job of this kind, by corroborating student and parent reports of college financing arrangements against college student aid records.

Our third desideratum for a longitudinal survey is a long span of years, both pre- and post-college. Owing to the problem of selection effects, good data on pre-college background and experience, extending even as far as early childhood, can be of enormous value. Following a cohort well into the post-college years is also of great value, since it is very plausible that the effects of college are long-lasting, and some may take a long time to manifest themselves. People often do not attend college immediately after high school, even those who graduate from college typically now take more than 4 years to complete, and many people obtain post-collegiate education. After education is complete, there is often a period of job experimentation and search that may last 3 to 5 years or more. For many people a long-term career profile does not begin to jell until they are in their early 30s.

These long time spans are obviously quite frustrating for two reasons. One is the perennial desire for prompt answers to urgent questions. The other is the worry that the world changes so fast that data obtained about the college experiences of people now in their 40s may be irrelevant to the educational experiences of those in school now. In practical terms, and given limited budgets, at any particular time this question comes down to two more focused choices: should we do another round of an "old" longitudinal survey, or should we use those resources to start a new one? And, should a new longitudinal survey start with early childhood, or pick up people at a point closer to maturity? Although the answer is always a matter of judgment, we would express a preference for the long view, based on the suspicion that the system will always tend to be biased in favor of a short-run view. Our reasoning is this: we really should view work in this area as "basic research." The social science community is far away from having reliable knowledge about the effects of college or of how those effects are brought about. One of the advantages of a deeper understanding would be an ability to explain how differences in the educational system influence educational outcomes from one decade to the next. Investing now in the data collection efforts that will eventually bear this fruit, and summoning the patience to await the maturing of those data sets, seems to us the more sensible course.

THE NEED FOR BETTER DATA ON EDUCATIONAL TREATMENTS

One of the major lessons of this review is the high potential payoff from data that get closer to the actual educational "treatment" than existing national data sets do. Unfortunately, another lesson is the great difficulty of getting such data in a reliable form, and at reasonable cost. The appeal of such data, as should be clear from our earlier discussion, is the opportunity they would provide to get a clearer fix on how educational efforts and resources get translated into impacts on students' learning and hence on their later lives. Such data would help overcome two major limitations on existing work on higher education based on national data sets. First, available data will often fail to detect what may be large differences in the educational treatments received by students. For example, two institutions may have identical levels of instructional spending per student, or of numbers of library volumes, while offering very different instructional or library experiences to their students. Although this fact will not introduce any econometric bias into studies that ignore these differences, it will reduce, perhaps substantially, the precision of any findings. Second, existing studies average over educational experiences that, quite likely, vary substantially across students in the same institution. Ignoring this variation will also reduce the precision of estimates. But, more significantly, unmeasured variation in the educational treatments applied to different students may be a source of bias. If differences in such student characteristics as social background, academic ability, or motivation influence the educational treatments those students receive, there will be a tendency to overestimate the impact of these background variables on educational results, and to underestimate the impact of educational resources.

Better measurement of educational "treatments" would be of great value in estimating educational production functions, in assessing the returns to different types of education, and in studying the cost-effectiveness of different educational strategies.

As mentioned above, in existing longitudinal studies information about the educational environment of the colleges attended by a student in the sample is provided principally by linking the survey data to Integrated Postsecondary Education Data System (IPEDS) data on the institution. IPEDS, an institutionwide survey, provides no information on differences in educational resources provided to different individuals in the same school, or even to different groups of students (such as graduate and undergraduate students) within the school. Moreover, even the information on the resources applied to the average student are limited. IPEDS, a finance and enrollment survey, does not describe physical inputs, but rather only the dollar amounts spent in broad categories. It is also difficult in some cases to distinguish dollars spent on educational purposes from dollars spent for other institutional purposes in the IPEDS data. Finally, one important educational input—the quality of other students—is not measured at all in the IPEDS data.

Improving this situation would be a great help in improving understanding of the costs and benefits of higher education, and especially in helping learn about the relative costs and benefits of different types of or approaches to higher education. Two rather different kinds of improvements in data on higher education inputs should be distinguished. First is better measurement of actual inputs, rather than simply dollars. Thus, data on class sizes, on instructional methods employed, on the role of graduate assistants versus faculty in teaching, and so on, could be enormously helpful. Ideally, one would have data individualized to students (such as the sizes of classes experienced by a given student in a longitudinal sample). More realistically, one might hope for such data by classes of students (freshmen, sophomores, and so on) But even to have such data for the average student in a school would be a real improvement.

The second type of data improvement would be more refined measures of costs. Thus, for example, it would be very helpful to be able to distinguish costs of graduate and undergraduate education in the IPEDS data. Refinements of some expenditure categories in the IPEDS survey would also be welcome—a favorite example is including the costs of the admissions office in student services.

Conceivably, some refinements of the latter sort might be introduced in future generations of the IPEDS survey. However, as a survey intended to be a census of all postsecondary institutions, it would be unreasonable to expect IPEDS to be a vehicle for collecting detailed data on educational treatments. Several strategies are offered here that may be worth considering to enable NCES to make progress on this front.

First, NCES might consider doing a "long-form" IPEDS for a sample of institutions, analogous to the Population Census long forms. For example, if 5 percent of postsecondary institutions were selected for more intensive treatment, that would amount to about 160 public and private not-for-profit institutions, and a somewhat larger number of proprietary institutions.

Ideally such a long form should be administered at random (as the Population Census does), but even if the institutions had to be selected on a voluntary basis, the effort might be worthwhile. It would also be reasonable for NCES to reimburse institutions for the expense of undertaking a more thorough study.

A variation on this idea would be to link an intensive effort to measure institutions' educational practices to the participation of those institutions in a longitudinal survey. It would be reasonable in such a framework to include fewer institutions in the study, with more students from each institution. The trade-off is that one would have less variation among institutions but more information about each one. If it were possible in the context of such a study to measure actual variation in the educational resources provided to different students, being able to include this kind of variation would probably more than make up for having fewer institutions in the sample.

Finally, NCES might consider ways of approaching getting these data through cooperation with institutions that are interested in doing such studies internally. Some institutions are interested in gathering detailed data on their internal educational practices, and in using those data to improve their practices. For the individual institution, the inability to make comparisons to other institutions is a real drawback. NCES might have some opportunity to help to support individual institutions in making such efforts, and might be able to help standardize the efforts of different institutions in order to facilitate comparisons. The loss of randomization implied by this strategy is a significant drawback, but the advantages of having institutions as enthusiastic partners in the effort would be considerable.

We offer all these suggestions tentatively. We recognize that any of these efforts would be expensive and would challenge a general reluctance of colleges and universities to make detailed information about their internal practices known. Yet the potential gains in understanding are considerable.

MAKING INTERNATIONAL COMPARISONS

More and more countries have been attempting to measure the performance and effectiveness of their higher education industries, giving rise to the obvious question of whether there are particular indicators that would be of use in making international comparisons. If so, it would be important to make sure that NCES data sets include such information.

A recent monograph (Gaither, Nedwek, and Neal 1994) reviews some of this literature, dividing performance indicators into three types: input measures (test scores and secondary school performance of entering students, prestige of programs from which faculty received Ph.D. degrees, and so on); process indicators (library use, meetings with faculty advisors, and so on); and output measures (number of degrees awarded, graduation rates, faculty publications, percentage of students going to graduate school, and so on).

In their discussion of performance indicators in Britain, they point out that most of the indicators are input rather than process or output measures. Key measures include admittance rates and entry scores for undergraduates, their subsequent graduation rates and postgraduate employment experiences, and for faculty, research grants and publications. In terms of cost measures, staff/student ratios, unit costs, and institutional revenue and expenditures data are all used, although the authors report that difficulties in cost allocation procedures have made it very hard to evaluate efficiency.

Indicator systems in Canada also center on simple input and outcome measures, with relatively little on the process side. Popular indicators include time to degree, degrees granted, and various expenditure types. Typical indicators used in Australia are graduation rates, class size, and a series of "destination outcomes," including post-graduation employment, study, and salary. Again, process measures are neglected. The Netherlands concentrates on such teaching indicators as the number of students and their length of study, while Finland relies on similar aggregate measures. Sweden also concentrates on basic student enrollment and graduation indicators. Denmark supplements this sort of data with "customer satisfaction" information gathered from interviews with students, graduates, and employers.

In summarizing the indicators used in the seven countries they examine, Gaither, Nedwek, and Neal conclude that certain simple input and output measures—with some variation—are commonly used by educators and government officials in a variety of contexts. This raises two questions: do existing data sets in the United States allow us to compute these measures; and are these measures really of use in comparing the costs and benefits of higher education across the world?

The answer to the first question is "yes." It is not very difficult to collect information on the number of degrees awarded or total educational expenditures. In fact, variables of this type were mentioned in our earlier discussion. But, for example, in an analysis of the cost effectiveness of public higher education expenditures in the United States, we would hesitate to simply divide the number of degrees by state spending and compare that "productivity" measure across states. There is no reason to expect that the educational quality is similar enough to give any real meaning to this ratio, and the same can certainly be said for comparisons across countries.

We are therefore rather skeptical that data could be developed that would permit meaningful international comparisons of the relative costs and benefits of America's postsecondary education enterprise. However, the recommendations we have made here concerning data collection and analysis could help us increase our understanding of the costs and benefits of higher education *within* our country. The payoff would be considerable, both from the standpoint of individual students and colleges and from the nation as a whole.

REFERENCES

Gaither, G., Nedwek, B., and Neal, J. 1994. *Measuring Up: The Promises and Pitfalls of Performance Indicators in Higher Education* (ASHE-ERIC Higher Education Report No. 5). Washington, D.C.: The George Washington University, Graduate School of Education and Human Development.

Special Issues in Postsecondary Education and Lifelong Learning

David W. Breneman
Fred J. Galloway

ABSTRACT

In an effort to improve the data collection abilities of NCES, this paper identifies six emerging research areas in postsecondary education and lifelong learning: institutional finance, postsecondary assessment, loans and student indebtedness, the school-to-work transition, technological change and distance learning, and the proprietary sector. For each of these emerging issues, we provide both a contextual discussion and a review of the extent to which existing NCES databases can respond to these emerging issues.

Recommendations are provided for both data collection and data dissemination activities and are ordered by our perception of where "the biggest bang for the NCES buck" might be. These include increases in the proposed coverage and sampling frame of several of the data sets; the establishment of agreements with outside agencies to provide information that was previously self-reported; the creation of a new database that surveys high school graduates each year; an increase in the frequency with which the National Postsecondary Student Aid Study (NPSAS) is administered; an increase in the number of analysis reports issued each year; and an increase in both the coverage and availability of the public access versions of several of the data sets.

INTRODUCTION

Few areas of social policy are devoid of turmoil and disagreement regarding future directions, and the world of postsecondary education is no exception. Indeed, during the first years of the 1990s, higher education funding from state governments was the one area of broad state responsibility that saw a percentage decline in support. As a consequence, tuition levels increased sharply, access for low-income students was reduced, and worries about college affordability increased for middle and even upper income families. Adding to the dilemma of families and students is the pivotal role of postsecondary education as the gateway to challenging and remunerative employment, coupled, however, with a growing dispersion of opportunities and earnings, even among the college-educated. As higher education becomes essential, its economic payoff appears more like a lottery, with big winners and losers. The recent explosion of debt financing adds further tension to this relationship between investment in college and economic return.

Broad social and economic developments such as the above are beyond the power of any data collection exercise to anticipate or influence; nonetheless, as the scale of costs and benefits to students and to society expands, it is incumbent on the National Center for Education Statistics (NCES) to monitor and help policymakers and others interpret trends in the industry. Many of the data sets currently collected on postsecondary education and its students perform that function effectively, but we believe that cost-effective improvements are possible. Our discussion is organized around six emerging research issues, described in the next section. Following a brief discussion of each issue, the paper examines the current state of data collection in each area. The paper then concludes with recommendations for modifications and enhancements of NCES data collection practices.

EMERGING RESEARCH ISSUES

Institutional Finance

As one considers the last four decades, the overriding picture of postsecondary education is of an expanding, growing industry, with increasing enrollments, growth in the number and size of institutions, employment, and resources. Only recently have these patterns begun to shift toward stasis, with a focus on retrenchment, doing more with less, and growth by substitution. No one knows for certain whether this recent trend will continue, but no key revenue source seems poised for sharp increase. As noted in the Introduction, state support has slowed, and federal dollars for student financial aid and for research are under similar budgetary stress. Tuition increases of recent years have slowed, as private colleges and universities fear that they are pricing themselves out of reach, while political reaction to public tuition increases has forced a slow down. Philanthropy appears to be the source most emphasized, as both private and public institutions step up their fund-raising efforts. Suffice it to say, however, that the resource outlook for most colleges and universities is as cloudy today as it ever has been.

Among the responses of colleges and universities, two will serve as examples of the sharp changes under way. In the private, nonprofit sector, institutions are engaging in calculated price discrimination in the form of student aid discounting to fill their classes and to attract students of particular quality. This tendency toward discounting has accelerated in recent years, as colleges literally fight, in some cases, for survival. The economics of discounting in this sector is only beginning to be understood, and the Integrated Postsecondary Education Data System (IPEDS) database is only partially adequate for analysis tasks. In particular, that database does not differentiate between types of discounting, and is not sufficient for monitoring institutional financial stress in a rapidly changing environment.

In the public sector, talk is increasing of privatization in some form, with state universities becoming state-assisted institutions, relying more on tuition and private fund raising than on state support. The extent to which this trend is occurring is a matter of conjecture, because databases are not clearly focused on such issues. In both this and the prior example, higher education could be better served by improvements in financial data collecting, and our suggestions for change are noted in the next section.

Assessment

As the private and social costs of postsecondary education have grown, it is not surprising that both families and policymakers have sought more information about the benefits of higher learning. We think of the "assessment movement" as a rational response to the need for better measures of educational outcomes, thereby permitting individuals and society to calculate rough cost-benefit ratios. As postsecondary education has evolved from an elite to a mass phenomenon, more sophisticated measures of educational results have become necessary, responding to the diversity of students and reasons for enrollment.

Economic rate of return calculations, first developed in the 1960s, helped to fuel the growth of college enrollment, as the basic message was that college was a good investment. Today, however, many question whether the country has moved too far, with growing numbers of students enrolled in remedial courses and not completing their degrees. Some argue that we have too limited a range of postsecondary learning options, pointing to German apprenticeships as a better model. And how does one evaluate the many students who begin but do not finish programs? Should such students be viewed as "wastage," with the focus turned to retention, or have they gained something of value, and is concern misplaced? These are among the important policy questions that NCES longitudinal data sets can help to answer, provided certain changes are made.

Loans and Student Indebtedness

One of the most dramatic shifts in college finance in recent years has been the growing share of economic costs borne by students, financed primarily through increased student loans. While much of the policy focus has been on aspects of loan repayment—default rates, income-contingent options, and so forth—more fundamental, long-term issues of human behavior are involved. High levels of student debt may affect career choice, marriage, and child-bearing decisions, as well as patterns of saving and consumption over the life cycle. While many have speculated about such issues, very little empirical information has been available to analysts seeking to understand these relationships more clearly. The growth of student debt appears unstoppable; thus, it behooves us to begin to collect data that can help us understand the long-term implications of this social choice.

School-to-Work Transition

As the labor market grows ever more complex, the old verities about high school transitions to work, or high school transitions to college and then to work, are increasingly inaccurate. High school graduates today face a limited, and for the most part, unappealing set of choices—dead end jobs, the military, unemployment, crime, or college. Not surprisingly, college appears to be the best choice, but then there is the issue of which college, which major, and at what price. After college graduation, the choice of work or graduate school comes up, and throughout one's career, the decision to return full or part time for further education is a continuing dilemma. In short, the worlds of formal education and of work are no longer clearly segmented by age or employment situation. This blurring of circumstance yields a need to know more about the choices facing people at several stages in life, and the realistic options open at each stage. Modifications of the several longitudinal files maintained by NCES is the obvious way to enhance our knowledge in this area.

Technological Change and Distance Learning

Perhaps the greatest imponderable in our current situation is the prospect of technology for transforming the way we deliver education. One can hear the voices of prophets proclaiming a new millennium, in which education as we have known it will diminish, even vanish, from the scene. No longer will physical places called "colleges" or "universities" be necessary because anyone will be able to tap the resources of electronic information systems and video texts. New suppliers are expected to enter the market, providing education at much lower cost because they are not freighted down with either the physical plant or the outmoded traditions of academia. In this view, all that saves colleges and universities is the near monopoly on credentials, a too fragile reed to survive the onslaught of technological advance.

Others see the new technology as the salvation of college and university education, because at last a way may be found to escape the "cost disease," the tendency for unit costs to rise annually by about 3 percent above inflation. And still others dismiss the talk about technological revolution as yet the latest over-promoted fad, analogous to the promises made in an earlier generation for educational television. Much rides on which vision is the accurate one, and information about trends is clearly crucial to the evaluation of claims and promises. NCES can play a key role in helping to shed light on this important, but vexing, topic.

Proprietary Institutions

Much of the terrain of traditional, non-profit higher education is well-mapped by the various NCES databases, but the burgeoning universe of profit-making schools and colleges is only lightly covered by existing surveys. These schools are the largely hidden world of postsecondary education, having grown dramatically in response to eligibility for federal grants and loans by their students. Claims and counterclaims about their effectiveness are lobbed back and forth by educators and policy analysts, reflecting largely newspaper accounts and anecdotal information rather than hard data. The simple fact is that we do not know how many are doing a good job and how many are simply exploiting our most vulnerable young people. NCES would do society a great service by focusing on this group of schools and developing a systematic data collection effort that could be effectively implemented.

CURRENT STATE OF DATA COLLECTION

Using the above six categories of emerging issues, this section of the paper discusses the extent to which existing NCES and other databases are able to respond to the questions raised in each area. This discussion leads, in turn, both to modest suggestions for incremental change and to a small number of recommendations for substantial new surveys.

Institutional Finance

In an effort to maintain enrollment levels and ensure a diverse student population, many institutions have dramatically increased their contribution to the student's financial aid package. Although this growth has occurred among all types of institutions, the biggest increase has been at private, 4-year colleges and universities that either use institutional aid to meet enrollment targets, or as a form of merit aid to attract academic stars. This leads to two sets of research questions: first, what effect does this increasing reliance on institutional aid have on institutional health; and second, how does the growth in institutional aid affect such student outcomes as enrollment and persistence?

Unfortunately, current NCES databases provide little if any help in addressing these research questions. For example, one important bit of information needed to address both research questions is the ability to distinguish among institutions that use institutional aid to meet enrollment targets versus those schools that use it solely to attract stars. Without the ability to discern institutional motive in the awarding of aid, it becomes increasingly difficult to apply the appropriate standard of institutional health. For example, for institutions seeking to diversify their student population, net tuition revenue (gross tuition revenue minus institutionally provided aid) seems an inappropriate metric, yet for those institutions trying to meet enrollment targets, it may well be the appropriate measure of institutional health.

Currently, IPEDS collects information on institutional characteristics, including enrollment and financial statistics. However, the categories used to collect the information provide only gross measures, nothing approaching the context required to differentiate institutional motive in the awarding of this particular kind of aid.[1] Even those measures that might provide some hint of administrative context get "scrubbed" by the state higher education associations before arriving at NCES, further reducing potentially interesting variation across institutions. When combined with the other well-known limitations of IPEDS, one wonders if this data set could be successfully reconfigured to address these issues, or if some "student aid management" survey needs to be created.

The limitations embedded in the IPEDS system also extend to the student-based research questions involving enrollment and persistence. Even if IPEDS allowed us to differentiate among institutional motive in the awarding of this type of aid, the measurement of student-based outcomes would most likely occur through NPSAS, where enough financial aid information is collected on individual students to effectively address the second research question. To do this, however, would require that IPEDS and NPSAS be linked, so that the characteristics of IPEDS institutions would be matched with individual students in NPSAS. Even with this linkage, however, the NPSAS sampling frame would probably need to be increased so that there would be enough students in the sample to provide an adequate statistical test for the various student-based propositions concerning institutional aid. And if the information is to be used for policy decisions, then it needs to be available in a timely manner, something that is currently unavailable within the 3-year cycle under which NPSAS operates.

Assessment

As more students return to school for just a few skill-specific courses, the quality of instruction and amount of learning that takes place in postsecondary classrooms becomes increasingly important. To address effectively the growing importance of assessment in postsecondary education, several important modifications must be made both in terms of the information collected and the way in which it is collected.

To understand the importance of the particular information that needs to be collected, it may be helpful to classify students into one of three groups: degree earners, those with some college experience, and those who just enroll in a few specific courses. Although the later group may be growing the fastest, each group faces its own unique assessment needs. For those with college degrees, typical measures of societal assessment include the degree itself, the school attended, annual earnings, cumulative grade point average (GPA), as well as scores on such standardized tests as the Scholastic Aptitude Test (SAT), the American College Test (ACT), and the Graduate Record Exam (GRE). Currently, most of this information is collected by NCES in NPSAS, Beginning Postsecondary Students (BPS), and Baccalaureate and Beyond (B&B), so little additional information is needed for this group of students.

For those individuals with some college experience, the most important assessment measure may be the number of credits earned, the school attended, and annual earnings. Fortunately, most of this information is also collected by NCES. However, for those students who just enroll in a few postsecondary courses, two important pieces of information need to be collected. The first, and perhaps most important, concerns the motivation of the returning student. It seems that if the individual is taking the course for entertainment or personal enrichment, it makes little sense to apply any tools of assessment to the student's performance in the class. However, if the returning student is taking the course for a job-related reason, then some sort of value-added assessment measure is appropriate. The selection of appropriate assessment measures for these students, however, is quite controversial. Short of requiring both pre- and post-tests for this group, traditional measures such as the grade in the course, the quality of the instructor, or any increase in earnings may have to suffice. To the extent that this information is not currently collected, it should be added to the student-based NCES data sets.

Perhaps even more important, however, is the reliability of the information currently being collected. Although much of the relevant assessment information collected in NPSAS is done through transcripts, some of the information is self-reported. As demonstrated by numerous researchers, such self-reported data as annual earnings, test scores, GPAs, and years of education tend to be significantly overstated, introducing enough measurement error into the variables to make them virtually unusable in any statistical analysis. Since this information is already being collected, it makes sense for NCES to go straight to the source wherever possible. For example, test score information could be gathered from the College Board, and income and earnings information from the Internal Revenue Service, rather than relying on any self-reporting. In fact, if such a match could be accomplished, fewer questions would have to be asked in the surveys, thereby freeing up additional resources either to expand the sampling frame or to solicit additional pertinent information.

Loans and Student Indebtedness

In the last few years, students have become increasingly reliant on loans to help finance their postsecondary education. Since increases in student indebtedness have implications for future patterns of domestic consumption, an emerging issue concerns both the extent of the problem (exactly what is the combined debt load of postsecondary graduates) and how these higher debt levels influence acquiring such major items as automobiles and new homes.

Although the lack of up-to-date information on indebtedness has been a major problem in fighting to save subsidized student loans during the recent federal budgetary debate, the timing of this information is more of an issue than its ultimate acquisition. As currently configured, information on student indebtedness is available for both undergraduates and graduate/professional students through NPSAS, and will be available in the future through BPS and B&B. However, the 3-year cycle that drives NPSAS means that to get information from the 1992–93 academic year, one needs to wait roughly 3 years. Given the rapidly changing nature of financial aid programs in this country, the 3-year cycle means that analysts are always behind the curve, forced to speculate on emerging patterns or to conduct their own surveys. Moving NPSAS to a 2-year cycle would help ameliorate this problem.

To a large extent, the same timing issues are relevant in the discussion of BPS and B&B. Originally designed to alternate with each other as a companion to NPSAS, these data sets contain important information on overall student debt levels, but suffer from the same long-cycle problems as NPSAS. Even more important, however, is the need for these surveys to follow students well into their careers, so that the full effects of their postsecondary financing decisions can be documented. Given the standard 10-year repayment period for most student loans, it would seem that individuals would need to be tracked for at least 10 years, and probably more, to capture the behavioral changes that occur as their student debt is finally retired. As such, we strongly advocate both shorter cycles and more follow-ups for the surveys to become truly effective tools for both researchers and policymakers.

School-to-Work Transition

Given the rapidly changing nature of work in this country, today's high school and college graduates face an uncertain future in terms of job availability and rapidly changing skill requirements. To help them plan for this transition, more information is needed on the career paths of recent graduates, as well as intertemporal changes in the distribution of job offers for recent graduates.

To address these issues, contemporary information is needed on two sets of graduates, high school and college. For college graduates (both undergraduate and graduate), the amount of information currently collected by NCES may have to be expanded to include more information on job offers, search strategies, and starting salaries, but the larger issue is the frequency with which the data are collected. Since most of the pertinent information is contained in B&B (which alternates with BPS in NPSAS), the resulting 6-year cycle provides information that is of little practical value to the ultimate consumers of such information—researchers and recent graduates.

Additional support for this proposition comes from both the National Academy of Sciences (NAS) and Tom Mortenson's "Postsecondary Education Opportunity Research Letter." As described in the NAS publication *Reshaping the Graduate Education of Scientists and Engineers*, members of the Committee on Science, Engineering, and Public Policy argue:

> Graduate scientists and engineers and their advisors should receive more up-to-date, accurate, and accessible information to make informed decisions about professional careers. We recommend that a national database on employment options and trends be established (National Academy of Sciences 1995).

Ironically, their recommendation comes at a time when the most reliable source of undergraduate starting salary information has recently been discontinued. In writing about the termination of the Endicott survey on starting salaries of college graduates, Tom Mortenson writes:

> Currently, several of the data sources that reveal the condition of educational opportunity in the United States are under assault. The Endicott survey data on starting salaries of college graduates . . . that was collected and reported by Northwestern University since 1947 was ended in 1994. A 48-year time series of data used in numerous econometric studies of student demand for education has been terminated (Mortenson 1995).

To provide this information in a more timely manner, we recommend that B&B be either included in every NPSAS survey, or that B&B continue to alternate with BPS, but that NPSAS be moved to a 2-year cycle. In this manner, the requisite information would be available every 3 years under our first option, or every 4 years if the second option were adopted.

While our recommendations for improving the timeliness of information on recent college graduates may be resolved by simply changing the cycle on which several databases operate, a more serious structural problem exists for high school graduates, the most overlooked group of individuals in the NCES sampling universe. Although information is collected every 3 years (through NPSAS) for those high school graduates who enroll in college, no information is collected on those who directly enter the work force. For these individuals, a national "black hole" currently exists in terms of up-to-date information on starting salaries and potential career paths. To generate this information, we recommend a short longitudinal study, conducted every year, of our nation's graduating seniors, with at least one 2-year follow-up survey. In this manner, contemporary salary and career information could be gathered and made available in a timely manner for this long-neglected group of individuals.

Creating such a national database would also provide a wealth of information on access and choice for those graduating seniors who elect to continue on to postsecondary education. Surprisingly, this information has been collected only four times in the last 35 years, in 1972, 1980, 1982, and 1992, through the National Longitudinal Study (NLS), High School and Beyond (HS&B), and the National Education Longitudinal Study (NELS) databases. Given that access and choice are two of the main reasons for the very existence of financial aid, it is truly shocking that this information is not collected regularly by NCES. In fact, if such a database were created, it could be linked up with the Common Core of Data (CCD), Schools and Staffing Survey

(SASS), and IPEDS data sets, so that from an information perspective, the entire transition from high school to college (including the characteristics of the high school, the college application process, and the college eventually selected) would be seamless.

Technological Change and Distance Learning

As information technologies continue to revolutionize the way individuals both work and learn, an increasing number of students will spend time in "nontraditional" classrooms. To evaluate the effectiveness of these new modes of teaching and learning, information must be gathered not only on the methods of delivery but also on a variety of student-based outcome measures.

In collecting this sort of information, there are several issues that NCES needs to address. The first involves from which end of the delivery system the data should be collected—the user or the institution. To provide overlapping coverage, we recommend that the data be collected at both ends. In this manner, questions could be added to NPSAS and BPS that measure the availability and frequency of this type of learning at the student level, and similar questions could be added to IPEDS and the National Survey of Postsecondary Faculty (NSOPF) at the faculty and institutional level. In this manner, emerging trends could be identified at both the provider and consumer level, instead of lumping them together into the less interesting "user" level.

Another important issue is the timing of the data collection. Although the NSOPF appears to be on at least a 5-year cycle, the annual nature of IPEDS makes it a useful vehicle for collecting and reporting this sort of information. At the student-level, however, the timing problems previously discussed with NPSAS and BPS are again relevant. To remedy these problems, we encourage NCES to move NPSAS to a 2-year cycle, thereby providing consumer-based information on distance learning in a timely manner.

Perhaps the most important issue, however, is the ability of NCES to go "where the action is"—in this case, the proprietary sector. Although distance learning is occurring across all institutional types, NCES must be able to gather information from this sector or risk misstating the extent of this emerging technological innovation. Unfortunately, the ability of NCES to adequately measure anything in this sector is relatively weak—due largely to the refusal of many schools in this sector to share any information for fear of increased federal regulation. Although no simple solution seems apparent, NCES must increase their coverage of this sector, or risk relying on student-based information to capture this emerging and important trend.

Proprietary Institutions

As described in the last section, the coverage of the proprietary sector by NCES must be increased if the effectiveness of these for-profit institutions is to be debated publicly. Although both IPEDS and NPSAS provide some sectoral coverage, the lower response rates typical of schools in this sector make statistical inference an increasingly difficult task. When combined with the large numbers of schools regularly entering and exiting, even the notion of a "steady state" in this sector becomes somewhat meaningless.

To address these issues, NCES needs to find a way to increase institutional participation among for-profit institutions. If such a method were devised, the institutional sampling frame in the NPSAS and IPEDS databases could be increased, and inferences regarding this sector made more robust. Furthermore, by matching these institutions with their Internal Revenue Service records, financial information could be taken directly from their tax records, effectively solving the "self-reporting" problem. In this manner, both the quality and quantity of data from the proprietary sector would be significantly improved.

RECOMMENDATIONS

In this section of the paper, our recommendations for NCES will be presented. They flow logically from the previous discussion, and are divided into two groups: those dealing with the data collection process itself, and those dealing with the dissemination of information derived from this process. Within each group, the recommendations are ordered by our perception of where the "biggest bang for the NCES buck" might be.

Data Collection

Recommendation #1: Add the following information to the IPEDS, NPSAS, NSOPF, and BPS databases:

Although many of the changes recommended here represent only marginal additions to existing NCES data sets, we believe that their value added greatly exceeds the costs of implementation. For example, the IPEDS database could be made more useful in at least three ways: by adding a set of contextual questions designed to determine institutional motive in the awarding of various types of aid; by including a set of questions designed to solicit information on technological change and distance learning; and by expanding the sampling frame to include more proprietary institutions. In a similar manner, the NPSAS database could be improved by also expanding its sampling frame (which would make a potential linkage between IPEDS and NPSAS even easier), and by including questions on technological change and distance learning. Finally, both the NSOPF and BPS data sets could also be expanded to include questions on technological change and distance learning.

Recommendation #2: Enter into an agreement with the Internal Revenue Service, the College Board, and Educational Testing Service to provide some of the information currently collected through NCES surveys.

Although establishing such a linkage might require a substantial expenditure of political capital, the rewards would be enormous. For starters, such previously self-reported information as income, earnings, and some scores on standardized tests would be made substantially more reliable. In addition to the obvious benefits for both the consumers and practitioners of educational research, this would also mean that fewer questions would be asked in NPSAS, BPS, and B&B, thereby freeing resources either to increase the sampling frame in these databases or to ask other policy-relevant questions. By any measure, such a linkage would increase both the reliability of the data and subsequent analyses, in addition to either cutting programmatic costs or increasing the scope of the overall coverage.

Recommendation #3: Create a new database that surveys high school graduates every year, with at least one 2-year follow-up survey.

Creating such a database would allow researchers to address intertemporal issues of access and persistence among those high school graduates applying for college, as well as provide salary and career information on those students who enter the work force directly. Since this information has been collected for only 4 years out of the last 35, it would help researchers identify emerging trends among high school graduates, and could help current high school students decide on an appropriate career path. The data set itself could be linked with the CCD, SASS, and IPEDS databases to provide maximum information for the educational researcher and could be relatively "short and sweet," limited to perhaps as few as 8,000 high school graduates annually.

Recommendation #4: Move NPSAS from its current 3-year cycle to a 2-year one.

If NPSAS were administered every 2 years instead of 3, the timeliness of the resulting information and analyses would be greatly improved. Given the dynamics of postsecondary finance, this information needs to be collected at least every 2 years if researchers and policymakers are to stay reasonably ahead of the curve. Furthermore, since BPS and B&B alternate with each administration of NPSAS, the timeliness of their information would also be improved.

Data Dissemination

Recommendation #1: Produce more Postsecondary Education Descriptive Analysis Reports (PEDAR) reports.

Although many NCES users have restricted-access versions of the NCES data sets and many more use the DAS table-generating software, the PEDAR reports have perhaps the widest usage among consumers of educational research. Currently, five of these reports are produced each year, with topics ranging from the packaging of institutional aid to minority representation in higher education. Since the selection process involves a dozen or so proposed topics, it makes sense to produce at least a couple more reports a year, given the dependence of the research community on the reports. Furthermore, if the scope and coverage of some of the NCES databases are increased, then this should naturally be accompanied by the increased dissemination of analyses.

Recommendation #2: More public access versions of NCES databases.

As described in the above recommendation, those individuals without a restricted-access version of a particular NCES database are forced to rely on the PEDAR reports or to use the DAS software. Since this software limits the user to simple crosstabs and correlation coefficients on a subset of the variables, the question arises as to how much information the public should be able to access. At the least, we think that there should be public access versions of all the main NCES data sets, and if time and money permit, these public access versions should allow analysis on as many variables as possible. In this manner, the data that NCES worked long and hard to gather and clean would be made available to as many researchers as possible.

APPENDIX

To help identify the databases referenced in this paper, the following descriptions are provided by the National Data Resource Center:

Schools and Staffing Survey (SASS)

The SASS is an integrated sample survey of public and private schools; school districts; and principals and teachers. SASS was first administered during the 1987–88 school year, and again in 1990–91 and 1993–94. It will be conducted again in 1997–98. SASS consists of eight questionnaires: Public and Private School Administrator; Public School; Public School Teacher; Public School District Teacher Demand/Shortage; and Teacher Follow-up Survey. The following questionnaires were added for the 1993–94 school year: Public and Private School Library Media Center; Public and Private School Library Media Specialist/Librarian; and Public and Private School Student.

National Survey of Postsecondary Faculty (NSOPF)

The NSOPF is a survey of faculty in postsecondary institutions. The survey was initially conducted during the 1987–88 school year and was repeated during the 1992–93 school year. It consists of the following surveys: Institutional, Faculty, and Department Chair.

Common Core of Data (CCD)

The CCD is a set of five surveys sent to state education departments to collect data about all U.S. public elementary and secondary schools, local education agencies, and state education agencies. CCD contains three categories of information: general descriptive information on schools and school districts; data on students and staff; and fiscal data. The descriptive information includes name, address, phone number, and type of locale; the data on students and staff include demographic characteristics; and the fiscal data cover revenues and current expenditures.

High School and Beyond (HS&B)

The HS&B describes the activities of seniors and sophomores as they progressed through high school, postsecondary education, and into the workplace. The data span from the years 1980 through 1992 and include parent, teacher, high school transcript, student financial aid records, and college transcripts, as well as student questionnaires.

National Postsecondary Student Aid Study (NPSAS)

The NPSAS describes all types of postsecondary enrollees, ranging from full- and part-time students who attend private, for-profit (proprietary) institutions to those in prestigious public universities. Administrative records, with exceptional detail concerning student financial aid, are coupled with student interviews and data from a subsample of parents. Data are available from academic years 1986–87, 1989–90, and 1992–93.

National Longitudinal Study (NLS)

The NLS describes the transition of young adults from high school through postsecondary education and the workplace. The data span from the years 1972 through 1986 and include college transcripts.

National Education Longitudinal Study (NELS)

Beginning with an 8th-grade cohort in 1988, NELS provides trend data about critical transitions young people experience as they develop, attend school, and embark on their careers. Data were collected from students and their parents, teachers, and high school principals and from existing school records such as high school transcripts. Cognitive tests (math, science, reading, and history) were administered during the base year (1988), first follow-up (1990), second follow-up (1992), and third follow-up (1994). All dropouts were retained in the study.

Integrated Postsecondary Education Data System (IPEDS)

The IPEDS surveys most postsecondary institutions, including universities and colleges, as well as institutions offering technical and vocational education beyond the high school level. IPEDS began in 1986, replacing the Higher Education General Education Information Survey (HEGIS), which began in 1966. The components of IPEDS include Institutional Characteristics; Fall Enrollment; Salaries; Tenure and Fringe Benefits of Full-Time Faculty; Financial Statistics; Staff; and Academic Libraries.

Baccalaureate and Beyond (B&B)

Formally known as the Survey of Recent College Graduates (RCG), B&B is designed to analyze the occupational outcomes and educational experiences of bachelor's and master's degree recipients who graduated from colleges and universities in the continental United States. The survey was taken during the 1985–86, 1989–90, and 1993–94 academic years.

Beginning Postsecondary Students (BPS)

The BPS followed first-time beginning students from the 1989–90 NPSAS. NPSAS:90 asked additional questions of students eligible for BPS concerning background and experiences related to completion of postsecondary education. The BPS:90/92 data further describe the experiences during and transitions through postsecondary education and into the labor force, as well as provide information about family formation. Transfers, persisters, stopouts/dropouts, and vocational completers were among those who completed interviews in the first follow-up conducted in 1992. In the second follow-up, conducted in 1994, many will have completed a bachelor's degree as well.

NOTES

1. In response to recommendations of the Financial Accounting Standards Board, a committee made up of members of NCES and NACUBO (the National Association of College and University Business Officers) is working on changes to college and university financial statements, which will go a long way toward meeting these objectives.

REFERENCES

Mortenson, T. October 1995. "Starting Salaries of College Graduates 1947 to 1995." *Postsecondary Education Opportunity.*

National Academy of Sciences, Committee on Science, Engineering, and Public Policy. 1995. "Reshaping the Graduate Education of Scientists and Engineers." Washington, D.C.: National Academy Press.

Discussant Comments

JAMIE MERISOTIS

Let me begin by saying that NCES deserves a great deal of credit for what it accomplishes. As an agency that has been plagued by chronic underfunding, which operates with the federal procurement albatross permanently affixed to its neck and has had to fend off occasional attempts to politicize the Center's agenda and data collection vehicles, I have tremendous respect for the content and the quality of the work that NCES does. This conference, with expert guidance from MPR Associates, Inc., is a good example of the foresight and professionalism exhibited by NCES. I am delighted to be here and am honored by the invitation to participate.

The task before us today is to explore issues related to the national collection of data regarding postsecondary education over the next 5 to 10 years. This suggests that we need to have some sense, at least from a national policy perspective, of what the most important issues will be. So in beginning my comments about the two excellent papers from David Breneman and Fred Galloway and from Michael McPherson and Morton Schapiro, I would like to attempt to predict what those key issues will be. Because of the limited time we have, I will focus on just those issues that concern the federal government's interest in and influence on national data collection in postsecondary education.

First, it seems clear to me that the *federal role* in higher education will be a prominent if not dominant topic of discussion in the next decade. Undergirding this discussion of the federal role will be the central question of who pays for and who benefits from investment in postsecondary education. The personal, social, and economic benefits of postsecondary education will need to be clearly delineated and understood in the policy world in order to constructively engage in this conversation. The federal government already has attempted to address this topic at the K–12 level with the National Education Goals effort. In higher education, I think we will be examining how or if the federal government should play a role in setting goals for higher education; how those goals should be measured; and what happens if those goals are not achieved. I also think that the federal role in defining or delineating the distinctions among collegiate education, remedial instruction, and work force training will be important components of this discussion.

Second, the *level of support* that the federal government should be providing to pay for higher education will also be an important topic. What is the appropriate level of investment in postsecondary education from the federal perspective? What should the relationship be between the federal and state investments in higher education? What linkages, if any, should there be between federal support levels and institutional pricing? These are the kinds of questions that are posed in both papers and will form the core of the debate about federal support levels in the next several years.

Third, the issue of *program integrity* also will be critical. By program integrity I do not mean the current Higher Education Act usage of that phrase, which seems to confuse concerns about fraud and abuse in federal programs with what are the desired educational outcomes of those programs. Instead, I mean that the integrity of what the programs are supposed to do to influence the educational attainment of students—ranging from access to program completion—will be discussed.

Fourth, the appropriate methods for *regulating* or *deregulating* the federal government's interactions with higher education also will be essential. This is the flip side of the program integrity issue, and is related to determining what aspects of federal regulation might be eliminated without negatively affecting the federal government's legitimate interest in stemming fraud and abuse. Because NCES does not play a direct role in program management, this issue will be put aside for the purposes of this discussion.

Thus, in analyzing these two excellent papers in relation to what will be the most prominent issues of policy discussion in the next several years, I think we have two complementary pieces: the McPherson and Schapiro paper provides a road map for tracking the costs and benefits of postsecondary education over the next several years, which is key to determining what the federal role should be; and the Breneman and Galloway paper provides us with the key stops along the road, thereby helping to define what information we will need in setting federal support levels, and how we can track program integrity by deciding which outcomes of postsecondary education should be measured.

With respect to the particulars of the two papers, I am most compelled by McPherson and Schapiro's clear arguments for good longitudinal data. As the paper carefully points out, longitudinal data provide benchmarks on student attributes in order to examine how college changes the attributes. That analysis is critical to the task of determining who benefits from postsecondary education, which in turn will shape how we define the federal role.

Two specific points contained in the McPherson and Schapiro paper deserve careful consideration. First, I very much agree with the idea of creating a "long-form" IPEDS survey to collect detailed data on educational treatments, for the very reasons described in the paper. Participating institutions must be compensated for this extra effort, however. Second, I share the authors' observations regarding the utility of NCES data for making valid comparisons of higher education internationally. In this age of global economic and social systems, the inability to make reasonably precise comparisons represents one of our greatest shortcomings in national data collection.

The Breneman and Galloway paper contains many excellent suggestions regarding information that should be collected but currently is not. At the same time, however, I am wary of adding to the NCES burden in the absence of new resources. Simply put, I don't believe NCES is capable of doing more with less—that is what they have already been doing for more than a decade.

If new resources are available, I believe that the authors' proposal for a longitudinal database of high school graduates is an excellent idea. This should be a priority in any environment where new resources are available, since this database would allow us to conduct the kinds of seamless analyses of access, persistence, and work force performance of college graduates that we have attempted in the past using multiple, often incompatible, data sets. Such a database would take us a long way toward deciding what is the appropriate level of support from the federal government, and in assessing the integrity of programs with respect to influencing the educational attainment of students.

The recommendation by Breneman and Galloway for more published reports is important. If I have a criticism, though, it is that the reports currently produced under NCES supervision are unnecessarily dull. When every report appears to use the same adjectives—taken, no doubt, from an approved list—and when every report is similarly formatted and printed, a kind of mind-numbing effect can sometimes occur. In my office, we argue about which NCES publication contains certain information. These disagreements often end in frustration, since it is virtually impossible to distinguish among them. ("I think it was the blue book" is a sure-fire way to frustrate an opponent in such arguments.)

The only priority that the authors have identified with which I do not agree concerns research on proprietary institutions. Having spent several years during the 1980s conducting such research, I share the authors' frustration about the lack of reliable, consistent data. Unfortunately, I believe that expanding the sampling frame of various NCES surveys would be a day late and a dollar short. Given that the concept of "institution" is about to be radically transformed as a result of technological changes in pedagogy and educational delivery, focusing on 1980s-era concerns about proprietary schools seems misplaced.

Overall, I believe these two papers provide us with a template for future data collection and analysis, and are extremely valuable in informing key policy discussions over the coming decade. I appreciate the opportunity to comment on the papers and urge NCES to take the authors' recommendations seriously.

JIM MCKENNEY

Historically, postsecondary data collection has focused on traditional college-age students and the structures and procedures of the traditional 4-year college/university environment. As American community colleges have evolved, those historical definitions of postsecondary education have been assessed as increasingly dysfunctional by the 2-year college sector. Yet those definitions persist with surprising tenacity, which is surprising, since the size and growth of the community college enterprise would seem to inherently require a more customized approach. The national network of community colleges today numbers approximately 1,100 institutions in every state. In 1992, these colleges enrolled 5.7 million credit students and conservatively another 5 million non-credit students. The colleges enroll 44 percent of the nation's undergraduates and 49 percent of all first-time freshmen. The average age of a community college student is 29, with females constituting 58 percent of the college enrollment. About 47 percent of all minorities in college attend community colleges, and more than half of higher education students with disabilities attend public community colleges.

It is with that perspective and skepticism that this discussant reviewed the papers by McPherson and Schapiro and Breneman and Galloway. One can almost appreciate that researchers may have chosen to cling to the traditional definitions out of sheer fear of the complexity of the 2-year college sector. By comparison, community colleges may appear to be the moral equivalent to the Balkans for many postsecondary researchers. There is no separate typology for community colleges. Thus, everything is lumped together, undercutting substantially the ability to make finite distinctions. Using the tradition-bound National Center for Education Statistics (NCES) data sources, McPherson and Schapiro assessed the usefulness of these data in tracking the cost/benefit of postsecondary education. These authors argued persuasively that a direct correlation between cost benefits of postsecondary education and classroom activities is impossible to frame—especially, using the national data sets as they are presently constituted.

Under the tight parameters of traditional social research, McPherson and Schapiro suggest that you cannot tease out all of the other possible behavioral explanations for post-graduate performance. They are right, of course. And, their suggestions for merging data sets and seeking voluntary research contributions from individual institutions are great ideas—ideas that would seem to have merit due to the ability of volunteers to drive research to greater detail and at no great cost to NCES. Again the researchers point out the difficulties that come with merging existing data sets—sets that were created for different purposes. Thus, voluntary contributions from institutions or state systems might provide a better picture of the connections between certain desirable causes and effects. We might not have a national picture, but we would have a limited one. This reviewer would only add that such an endeavor should not be attempted without a substantial effort to enlist a respectable sampling of community colleges. Such states as Florida, California, North Carolina, and Illinois have historically maintained extensive data on their 2-year college systems.

A final point needs to be made about the McPherson and Schapiro paper and the concern regarding the uneven fit of traditional research approaches to higher education. During most of the discussion regarding the economic benefits of higher education, one was left with the impression that the researchers had in mind traditional liberal arts majors. What about the measurement of economic benefits that might correspond to graduates of occupational and technical programs at community colleges? For that matter, one could ask the same question about graduates of the professional programs at the university level. One would think that there would be a great payoff in looking at these questions with engineers, nurses, electronics technicians, and accountants. Also, while the authors speak of merging IPEDS and NPSAS, they have given no thought to the potential use of the National Assessment of Vocational Education (NAVE). Again, the railroad tracks might not make an even match, but these data are aimed at assessing occupational education at the secondary and postsecondary level. There just may be some value in looking beyond the traditional college student when it comes to seeking correlations between causes and effects.

Breneman and Galloway have attempted to improve the data collection abilities of NCES by focusing attention on the following six issues: institutional finance, postsecondary assessment, loans and student indebtedness, the school-to-work transition, technological change and distance learning, and the proprietary sector. Again, we have a very compelling argument for new ways of looking at data with an eye to cost effectiveness. For example, the suggestion is made again that IPEDS and NPSAS be connected and that NCES attempt a shorter 2-year cycle.

The authors make the case that institutions of higher education are under substantial financial stress with growth essentially being flat. There is the stated concern that this circumstance has led private institutions to engage in price discrimination through student aid discounting. Thus, private institutions have found a way through public student financial aid to defray their escalating costs in a flat market. This may be okay, but it is a public policy issue that can only be massaged with the existence of confirming data. Alternately, the case is made that financial duress has led the public sector to move toward varying forms of privatization, such as heavier reliance on tuition and private fund raising. It would seem to this reader that this is such a fundamental issue with respect to the true intent of financial aid use and misuse that NCES could hardly ignore the challenge. Heretofore, all of the attention has been focused on student abuse of aid, but the authors are raising a more subtle, but equally important issue.

On the other hand, it is doubtful that financial aid will be a good gauge for financial stress in community colleges. This is not to say that this sector is beyond economic duress. Rather, low tuition and high numbers of part-time students will mean that student aid will be a less robust intervening variable. Community colleges react to economic stress by lowering the full-time/part-time faculty ratio, cutting marginal curriculum/courses, seeking local bond market relief, and seeking infrequent and modest tuition increases. Of course, some community college students will seek recourse in some form of financial aid. The more likely student reaction will be reduced course loads, increased working hours, and the extension of years in college. Hence, the community college and the non-traditional student behave in very different ways from their 4-year counterparts.

The argument is proffered that there is a need for better outcome measures in order to permit individuals and society the ability to calculate rough cost-benefit ratios in making educational selections. The point is made that NCES does a good job in collecting data on degree earners and those with some college experience. However, the data are inadequate for those students enrolling for just a few postsecondary courses. The authors correctly point out that those courses taken for job-related reasons do have a value-added component that is not presently captured. From the standpoint of community colleges, Breneman and Galloway are beginning to spotlight a very large area of concerns surrounding the mapping of community college impact—the tracking of the growing number of non-traditional students, most of them adults, that drop in and out of college as if 2-year institutions were convenience stores. These students, some already having a degree, attend the college for the purpose of attending one or a series of classes in order to achieve a particular skill. Some of these students may achieve a degree over time, but rarely in 2 years and many never intend to earn a degree. Yet, they are there using these institutions in a value-added manner. Hence, community colleges and the night programs at 4-year institutions are becoming as important to the burgeoning number of adult students as they are to traditional college-age students. It seems that NCES must find a better way to map this terrain or forego the ability to comment with authority on this major growth sector of higher education.

Breneman and Galloway make a strong plea that the rising costs of higher education and the concomitant rise in loans and student indebtedness have implications for the larger economy and for student choices. The concern is raised that increasing levels of educational debt mean that students will likely defer other lifetime purchases and that they may, in fact, alter their occupational choices or lifetime goals as a result of debt incurred as students. It is pointed out that the data issue here is one of timing rather than that of an information vacuum. On the other hand, the authors suggest that the rapid changing nature of work calls for the development of more precise information on career path selection and on follow-up data with both college graduates and high school students moving directly into the work world. In short, the nature of the school-to-work transition has become the subject of such increased concern that NCES should not ignore the need to enhance the database for this purpose.

The point is valid as far as the authors take it. Community college professionals would point out the additional need to track the work-to-school behavior of adults. For these students, starting salary is less relevant than salary increases or job and occupation movement. For that matter, job retention may depend upon the acquisition of a new set of skills. Breneman and Galloway are correct in suggesting better follow-up data for working high school students and college graduates. But, the major story in higher education may be the work-to-school transition of the 25- to 40-year-old cohort.

Breneman and Galloway raise the specter that technological change and distance learning loom on the horizon with major implications for the delivery of instruction, the quality of instruction, and the financing of education. Ironically, it is the speed with which technology is influencing our world and the rapid response of consumers that raises questions about policy making that is dependent on NCES data collection. As stated, it is true that there is a need to assess quality and effectiveness among new modes of delivery, but this reader had the sinking feeling that we were all watching an avalanche in process and no one was sure what to do to avoid being run over.

Finally, the issue of the paucity of data surrounding the "burgeoning" world of the for-profit colleges is of major concern. The authors suggest that these institutions owe much of their financial success to the existence of federal student financial aid, but that these same institutions are not always very forthcoming with the requested information on their effectiveness. As in their earlier point about student aid discounting and privatization, Breneman and Galloway have raised another major policy issue with respect to the complex web of growing interdependence between financial aid and higher education. Federal policymakers cannot begin to address this issue effectively without better data from NCES. It would seem to this reader that this ought to receive the highest priority from NCES as student financial aid is the major federal investment in higher education.

The authors are to be congratulated for their penetrating look at the data sets and their suggestions regarding the applicability of these data to future issues in higher education. Clearly, the issues surrounding student financial aid are critical given budgetary constraints. Moreover, it appears that both papers call for a merging of data sets and a more user friendly timing of data access. All researchers were mindful that the desire to measure the benefit of higher education must be contrasted to the finite resources of NCES. Thus, most suggestions were made with an eye toward activities that sought economies of scale as well as new data yields. This reviewer thinks that NCES has received excellent suggestions from both sets of researchers. The major caveat is a reservation about the applicability of data generalizations to community colleges.

PAULA KNEPPER

I would like to thank the authors of these two papers for providing very thoughtful and complementary perspectives on improving NCES data in the area of postsecondary education. Mike McPherson and Morty Schapiro have presented a very thoughtful and expansive view of the need for longitudinal data at the postsecondary level and beyond. They have suggested several areas in which more information is needed about the college experience itself. They point out that it is necessary to illuminate the "black box" experience in order to more accurately relate education to outcome measures. However, their primary emphasis has been on the need for longer studies of any single cohort, and on the need for "good" longitudinal data.

Similarly, Dave Breneman and Fred Galloway have pointed out six specific areas where additional information is needed concerning postsecondary education. As in the McPherson and Schapiro paper, many of their data needs can only be met with additional longitudinal information. Breneman and Galloway have also provided a set of recommendations on how to achieve much of what is needed with limited resources.

As was mentioned yesterday, education is a very complex process at the K–12 level. Postsecondary education is even more complex; although it serves fewer people, it provides many more diverse experiences and serves a much more diverse population in terms of age and past experiences. Many people continue directly from high school and simply see it as more schooling. These are typically thought of as traditional students. But others continue after a hiatus from education only when they have perceived the need for additional education for a variety of reasons, not the least of which is to enhance their ability to acquire a better job. Some return because they want additional education, though not directly tied to obtaining a specific job. Still others do not complete degrees in the traditional order. For instance, they may return after completion of a bachelor's degree for vocational training of some type, often at the local community college or a private trade school. Others complete a second bachelor's degree instead of, before, or even after completing a master's degree or higher. And these non-traditional students are increasing in number.

Postsecondary education itself also has a split personality of sorts—vocational schools emphasize getting the skills required for a specific job, while collegiate education emphasizes expansion of knowledge not directly tied to a specific job. Galloway pointed this out in his presentation, and it was further emphasized by Jim McKenney in his discussion. The majority of postsecondary students attend either a private trade school or a community college sometime in their educational careers, but not necessarily as the first institution as is so often thought. As Breneman and Galloway have also pointed out, the transition from education and work is no longer neat and clean. McPherson and Schapiro further complicate the picture by pointing out that people in postsecondary education are for the most part there voluntarily, not because the state requires that they attend until a certain age.

These non-traditional patterns are not new—early in my professional career, I hired a programmer who had just completed work for a computer programming certificate at a proprietary vocational school in Northern Virginia. The previous spring, he had completed a bachelor's degree magna cum laude in psychology at a prominent state university, but had not found job prospects particularly promising. Several years after that, he started work on a master's degree in computer science, and when that was finished, he moved on to become one of the chief developers of the computerized database system used by one of the national grocery chains. This clearly was not the traditional path through postsecondary education, even though he started college right after high school. As a statistics agency, we cannot ignore these different paths through education and work.

We have been encouraged to think in broad terms without regard to money or other constraints. Both of these papers stress the importance of long-term longitudinal data, and both recommend that there be more longitudinal surveys with more frequent re-interviews, and that these be conducted over longer periods of time and include more subjects. But I think our real challenge is to consider the broad data needs and how we might begin to meet them within realistic resources. The suggestion to follow multiple high school or earlier cohorts more often and further through all of the possible education paths is unrealistic—sample size alone would be prohibitive when you think how many 8th graders, for example, you would need to sample to ensure that you had a representative sample of people taking each of the diverse paths through high school and later into and through postsecondary education, some as far as a Ph.D. or similar degree. Even in High School and Beyond, the numbers are too small for reliable analysis even into, much less through, the Ph.D. levels. Thus, it becomes clear that there are two real challenges to NCES:

1) We must keep in mind that even though we think of education as a continuum, in reality, K–12 is very different from postsecondary education in terms of both the "black box" process and in terms of goals and purposes. These differences must be reflected in the data we try to collect, how we collect them, and how we evaluate these data. Completing one level of postsecondary education no longer leads just to either the work force or the next higher level on the education continuum.
2) We must find ways to do more with less. This includes finding ways to reduce the time lag between data collection and data availability, while at the same time ensuring accuracy and completeness.

In order to enhance our surveys, we need to continually be aware of the changes that are rapidly occurring, several of which have been pointed out specifically by Breneman and Galloway—e.g., distance learning, increasing use of and capabilities of PCs, the ever faster expanding knowledge base and related curriculum concerns of what to teach in the time available. New issues are emerging almost daily. Because of these rapid changes, as several speakers pointed out yesterday, information is now also needed by others more quickly if it is to be useful. As a statistical agency, NCES provides data. But we need to make sure we do so in a manner that is consistent enough to evaluate change over time, yet is flexible enough to include new emerging areas and to provide information on their impact.

One way to do this is to be more imaginative in the use of technology. We saw a short demonstration yesterday of how the Web could be used. While this has obviously been put into effect in limited areas, how many of us have thought about its use in these or similar terms?

It has been suggested that we link into databases such as IRS for both student and institutional financial information. This has considerable appeal and could greatly reduce burden and cost to NCES. In this case, the link would be relatively easy; although the legal hurdles are higher, they are not insurmountable. A minor change in the laws governing IRS release of data and interagency cooperation could make this doable at reasonable cost. However, again we cannot ignore reality—the major impetus to change these laws will have to come from outside of NCES. But doing so would free up time and resources for other data collection and dissemination efforts. A similar case can be made for linkage with other databases, and in fact we do link to IPEDS for institutional information, and to ED Student Aid records for student aid and loan information. As other relevant databases are identified, the feasibility of linkages for data collection efficiency and accuracy should be investigated and implemented as appropriate.

In the area of longitudinal data collection, two seemingly opposing strategies have been suggested: fewer cohorts over longer time periods, and more frequent and overlapping cohorts. Both suggest more frequent re-survey intervals. This later point is the key to obtaining what McPherson and Schapiro refer to as "good" data. As they indicate, longer term surveys provide what no other type of surveys can, an indication of the impacts of experiences over longer time periods. However, given the reality of constant change and the non-homogeneity of postsecondary education, this approach can lead to misinformation as well as "old" information not seen as useful by the time it is available (years after the actual experience of interest). Part of the problem is the constricted sample of postsecondary attenders (a single age cohort rather than the full age mix of postsecondary attenders). The other problem with a longer term study starting in or before high school is that it cannot provide the sample size necessary for accurate evaluation of the various postsecondary experiences, a problem exacerbated by the continual reduction in participation at each higher level. However, this type of survey does help to tie the pieces of more segmented surveys together (as suggested by Breneman and Galloway).

Overlapping surveys, as recommended by Breneman and Galloway, while not providing long-term background information about individuals, would include all types of students at each level. However, this puts much more of a burden on NCES to develop sample designs that allow linkages between the unique surveys. For instance, a relatively small high school graduate sample with a 2-year followup could provide good access and choice information for immediate entrants, but would be too small a sample for postsecondary progress, completion, and postsecondary outcome information. However, a coordinated beginning postsecondary student survey such as BPS includes recent high school graduates as well as late entrants, and provides a full range of undergraduates (both traditional and non-traditional) and information on the undergraduate education. Again, however, too few will continue on to postbaccalaureate education to provide good information concerning education and outcomes at that level. Therefore, a new sample of only recent degree completers, as in Baccalaureate and Beyond, is necessary to provide a sufficient number of students who continue their education in order to obtain information about their experiences and outcomes.

For a system such as this to be useful, however, there needs to be a continuing commitment to conduct these various surveys on an appropriate schedule that, in fact, allows these comparisons between overlapping data collections. In addition, these overlapping surveys need to continue for sufficient time to make outcome comparisons, as McPherson and Schapiro suggest. However, at the frequency recommended by Breneman and Galloway, the data at each stage would be recent enough to not be considered "old," and the more frequent comparison cohorts would provide useful information concerning change.

Also, Breneman and Galloway have made it clear once again that NPSAS is vital and should be more, not less, frequent if it is to be useful to policymakers. They also recommend that the number of students within an institution be increased so that the data could provide useful information at the institution level as well as at the sector and national levels. Currently, only very gross statistics (such as the percentage receiving student aid or a student distribution by family income) can be calculated at the institution level, and not at all institutions. (NCES standards require each calculation be based on at least 30 individuals.) To be able to accurately calculate something like aid packages by class level within an institution would indeed result in a significant increase in student sample size. However, if structured properly, it could also allow both a BPS and a B&B cohort off of the same base-year survey, which they also recommended.

The larger samples that would be required by the recommendations of both sets of authors may not be as onerous as they appear on first blush. With current advances in technology, institutional computer assisted data entry (CADE), department record merges, and so on, this should become easier and less costly. In the same vein, I hope that the next Postsecondary Education Transcript Study (PETS) can be done more electronically than has been the case for the previous ones, and as a result will be more thorough and less costly.

To summarize briefly, these two papers have given NCES a great deal of guidance for the postsecondary longitudinal studies program. Though we have been working toward their suggested goals to some degree, they do provide additional support and guidance in terms of importance, frequency, size, length, content, and use. It is up to us to keep these goals in mind as we refine our data collection activities, and to creatively determine ways to make this set of surveys as useful and timely as has been suggested in these papers.

7

New Data Collection Methodologies, Part I: Observational Strategies

Large-Scale Video Surveys for the Study of Classroom Processes

---

James W. Stigler

INTRODUCTION

In thinking about what kinds of indicators NCES might employ in the next 10 years it is useful to consider the kinds of information that might be important to improve education. NCES might collect three broad classes of information: 1) data on outcomes, whether related to achievement, attainment, or other goals; 2) data on policy implementation, i.e., data that indicate whether or not educational policies have been implemented, and where implemented how effective the policies are; and 3) data relevant to the processes that produce educational outcomes.

All three types of data are important for the improvement of student learning and achievement. However, it is my view that too much emphasis has been placed on the measurement of outcomes, and not enough on the study of processes that cause the outcomes. The critique that W. Edwards Deming leveled at American industry applies just as well to American education: quality cannot be improved simply by mass inspection of products. Instead, it is necessary to reflect on the processes that produce quality products, and then take measures to bring those processes under control. Likewise in education, we cannot improve student learning simply by measuring outcomes; we must investigate the processes that lead to high student achievement.

Chief among the processes that cause student achievement must surely be the processes of teaching and learning that transpire inside classrooms. Yet, surprisingly, we collect virtually no data—whether at the national, state, or local levels—that yield information about what is going on in classrooms. This is not because such data are deemed unimportant: in a series of papers commissioned by the National Center for Education Statistics (NCES) in 1985, papers designed to set the agency's priorities for the next 10 years, the need for classroom process indicators was raised numerous times (Hall, Jaeger, Kearney, and Wiley 1985). Cronin (1985), for example, expressed concern with the paucity of data that could document curricular breadth or the actual implementation of curricular reform in the classroom. Moreover, Peterson (1985) cited a near complete lack of data on the quality of educational activities in the nation's classrooms, or even on the time teachers devote to various instructional activities. Including such indicators in the future was a clear recommendation of the 1985 report.

Ten years later, such indicators are still deemed important, but they are still lacking. A new NCES survey of leading educators and researchers, conducted by MPR Associates in the summer of 1994, again finds that the most frequently cited area in which better national data are needed is that of instructional practice. Yet the NCES *Condition of Education 1994* shows virtually no information at all concerning what happens in classrooms.

Probably the main reason for the continued lack of classroom process indicators is that what happens in classrooms is very difficult to describe and measure, especially on a large scale. What measures we do have are largely based on questionnaires in which teachers report on what happens in their own classrooms. Yet using questionnaires to measure classroom processes is problematic, as will be discussed below. Observation, on the other hand, would seem the natural way to study classroom processes. But observation is notoriously difficult and labor intensive.

Overview of This Paper

The first section of this paper will present a plea for the development of observational indicators of classroom process. The discussion will focus on what can be learned from observation, and argue for the advantages of video over live observers. The next section will explain some of the methodological issues that arise when video is used on a large scale. The final section of the paper will discuss the TIMSS Videotape Classroom Study, which I believe is the first attempt to use video for studying nationally representative samples of classroom teachers. This description will be detailed because the study really is the first of its kind, and much of what we have learned in this study will be helpful to those who follow. The software system we have developed for use on the project will also be described here.

Most researchers, on hearing the word "video," imagine a small-scale qualitative study. What I hope to demonstrate is the promise of using video for large-scale studies in which qualitative information can be easily combined with quantitative indicators.

WHAT WE CAN LEARN FROM CLASSROOM OBSERVATIONS

Having decided to study the processes of teaching and learning that go on inside classrooms, we must next decide how best to study these processes. In this section, a case will be made for using classroom observations, first by outlining the disadvantages of traditional questionnaire measures, and then by discussing the kinds of information that can be collected in observational studies. The focus here will be on two broad goals we might have for observational studies: first, to develop empirically validated models of instructional quality together with indicators for assessing instructional quality; and second, to monitor the implementation and effectiveness of educational policies.

Limitations of Questionnaires for Studying Classroom Processes

Most attempts to measure classroom processes on a large scale have used teacher questionnaires. Teachers have been asked, for example, to report on the percentage of time they spend in lecture versus discussion, the degree to which problem solving is a focus in their mathematics classrooms, and so on.

There are at least three major limitations imposed by the use of questionnaires to study classroom instruction. First, the words researchers use to describe the complexities of classroom instruction may not be used in the same way by teachers, or in a consistent way among different teachers. The phrase "problem solving" is a good example. Many reformers of mathematics education call for problem solving to become the focus of the lesson. But different teachers interpret this phrase in different ways. For instance, one teacher may believe that working on word problems is synonymous with problem solving, even if the problems are so simple that students can solve one in 15 seconds. Another teacher may believe that a problem that can be solved in less than a full class period is not a real problem but only an exercise. This kind of inconsistency is the rule in this country, where teachers have few opportunities to observe or be observed by other teachers in the classroom. Because teacher training in the United States generally does not engage teachers in discussions of classroom instruction, and because they are often isolated from one another by the conditions under which they work, teachers do not develop shared referents for the words used to describe instruction. Thus, although teachers may fill in questionnaires about their teaching practices, interpreting their responses is problematic.

A second problem with relying on questionnaire-based indicators of instruction concerns their accuracy. Even if teachers do interpret a question consistently, they may be inaccurate in reporting on processes that are probably at least in part outside of their awareness. Teaching is part planning, part performance. Teachers may be accurate reporters of what they planned for a lesson (e.g., what kind of demonstration they used to introduce the lesson), but they may be inaccurate when asked to report on actual aspects of teaching. Teachers process enormous quantities of information during a typical lesson and must continually adapt to changing circumstances, a process that happens too quickly to be under the teacher's conscious control. Observational studies of gender bias in teachers' questioning generally surprise teachers with their results: teachers who call on boys more frequently than girls, for example, have no idea that this is happening. Obviously, they would not be able to identify such a bias on a questionnaire.

A third limitation of questionnaires is their static nature. Teachers can only answer the questions we as researchers were clever enough to ask. Where an observer might notice something significant just by being in the classroom, questionnaires could not lead to the generation of new ideas or hypotheses in the same way.

Developing and Assessing Models of Instructional Quality

Developing observational indicators of classroom processes could serve two primary purposes: first, to aid in developing models of instructional quality; and second, to monitor and evaluate the effectiveness of educational policies.

Classroom instruction is a complex and multidimensional process. Nevertheless, we must have theoretical and methodological tools for studying classroom instruction if we are to improve it. Observational studies make it possible to develop indicators of classroom instruction that can then be used to develop and validate models of instructional quality. If this effort is to succeed, a number of indicators must be combined: we must examine the content of classroom lessons (the so-called implemented curriculum) as well as the methods teachers use to engage students in the content. That is, we must be able to examine the planned/structural aspects of instruction as well as the on-line implementation of instruction that occurs as the lesson unfolds. Evolving models of instructional quality will be linked to improved indicators for assessing instructional quality.

Monitoring and Evaluating Educational Policies

Once consensus emerges on classroom-based definitions of quality instruction, policies designed to improve the quality of instruction will emerge based on these definitions. Another role of observational studies, therefore, will be to monitor the implementation of these policies in classrooms, and to assess their effectiveness.

Policies designed to improve instructional quality will be similar to opportunity-to-learn (OTL) standards. As described by Porter (1995), these standards will offer two distinct advantages over outcome-based standards alone: 1) they can provide a vision of what good practice looks like; and 2) they can provide a system of school process indicators related to OTL goals.

A good example of these new policies is contained in the NCTM Standards, which represent a consensus on what high-quality instruction should look like in the classroom. Operationalizing this consensus in a system of classroom-based observational indicators will allow us to assess the degree to which the standards are being implemented, and to empirically assess the effectiveness of the teaching practices described in them.

ADVANTAGES OF VIDEO OVER LIVE OBSERVATION

Video has distinct advantages over live observation in the study of classroom processes. The next section will present these advantages.

Enables Study of Complex Processes

Classrooms are complex environments, and instruction is a complex process. Live observers are necessarily limited in what they can observe, and this, in turn, limits the kinds of assessments they can do. With video, the problem of "bandwidth" becomes manageable: observers can code video in multiple passes, coding different dimensions of classroom process on each pass. On one pass, for example, they might code the ways materials are used, on another the behavior of students.

Not only can coding be done in passes but it also can be done in slow motion. With video, for example, it is possible to transcribe the language of the classroom, enabling far more sophisticated analysis of complex discourse processes. Detailed coding of classroom discourse would be unthinkable without the capacity to slow down and listen again.

Increases Inter-Rater Reliability, Decreases Training Problems

Video also resolves problems of inter-rater reliability that are difficult to resolve in the context of live observations. Although it is possible to send observers out in pairs for the purpose of assessing reliability of indicators, it is often very inconvenient to do so. For example, if a study is being performed cross-culturally, or in geographically distant locations, it is often necessary to hire local observers. Bringing these observers together to check reliability is not usually feasible.

Having video also makes it far easier to train observers. With video, inter-rater reliability can be assessed not only between pairs of observers but also between all observers and an expert "standard" observer. Disagreements can be resolved based on re-viewing the video, making such disagreements into a valuable training opportunity. And, the same segments of video can be used for training all observers, increasing the chances that coders will use categories in comparable ways.

Amenable to Post-Hoc Coding, Secondary Analysis

Most survey data sets lose their interest over time. Researchers decide what questions to ask, and how to categorize responses, based on theories that are prevalent at a given time. Video data, because they are "pre-quantitative," can be re-coded and analyzed as theories change over time, giving these data a longer shelf life than other kinds. Researchers in the future may code videotapes of today for purposes completely different than those for which the tapes were originally collected.

Amenable to Coding from Multiple Perspectives

For similar reasons, video data are especially suited for coding from multiple disciplinary perspectives. Tapes of mathematics classes in different countries, for example, might be independently coded by psychologists, anthropologists, mathematicians, and educators. Not only is this cost effective but also it facilitates valuable communication across disciplines. The most fruitful interdisciplinary discussions result when researchers from diverse backgrounds compare analyses based on a common, concrete referent.

Merge Qualitative and Quantitative Information

Video makes it possible to merge qualitative and quantitative analyses in a way not possible with other kinds of data. With live-observer coding schemes the qualitative and quantitative analyses are done sequentially: initial qualitative analyses lead to the construction of the coding scheme, and implementation of the coding scheme leads to a re-evaluation of the qualitative analysis.

When video is available, it is possible to move much more quickly between the two modes of analysis. Once a code is applied, the researcher can go back and look more closely at the video segments that have been categorized together. This kind of focused observation makes it possible to see, for example, that the segments differ from each other in some significant way, and this difference may form the basis for a new code.

It also is possible with video to use example segments in reporting the results of the research. This gives the consumers of the information a richer qualitative sense of what each category in the coding system means.

Video Provides Referents for Teachers' Descriptions

Mentioned earlier was the problem that teachers lack a set of shared referents for the words they use to describe classroom instruction. Video, in the long run, can provide teachers, as potential consumers of the research, with a set of such referents. Definitions of instructional quality and the indicators developed to assess instructional quality can be linked to a library of video examples that teachers can use in the course of their professional development. In the long run, a shared set of referents can lead to the development of more efficient and valid questionnaire-based indicators of instructional quality.

A Source of New Ideas

A final advantage of video over other kinds of data is that it becomes a source of new ideas on how to teach. Because these new ideas are concrete and grounded in practice, they are potentially immediately useful for teachers. Questionnaires and coding schemes can help us to spot trends and relationships, but they cannot uncover a new way of teaching the Pythagorean Theorem. Video, especially if collected on a large scale, can be a treasure chest of such ideas.

ISSUES IN VIDEO RESEARCH

The next section will cover a number of issues that must be resolved in order to conduct meaningful video research.

Standardization of Camera Procedures

Left to their own devices, different videographers will photograph the same classroom lesson in different ways. One may focus in on individual students, while another may shoot wide shots in order to give the broadest possible picture of what is happening in the classroom. Yet another might focus on the teacher or on the blackboard. Because the intention is to study classroom instruction, not the videographers' camera habits, it is important to develop standardized procedures for using the camera, and then to carefully train videographers to follow these procedures.

The Problem of Observer Effects

Given that the camera is used in a consistent way, we must next consider the possible effect the camera might have on what happens in the classroom. Will students and teachers behave in typical fashion with the camera present, or will we get a view that is biased in some way? Might a teacher, knowing that she or he is to be videotaped, even prepare a special lesson just for the occasion that is unrepresentative of normal practices?

This problem is not unique to video studies. Questionnaires have the same potential for bias: teachers' questionnaire responses, as well as their behavior, may be biased toward cultural norms. On the other hand, it may actually be easier to gauge the degree of bias in video studies than in questionnaire studies. Teachers who try to alter their behavior for the videotaping will likely show some evidence that this is the case. Students, for example, may look puzzled or may not be able to follow routines that are clearly new for them.

It also should be noted that changing the way a teacher teaches is notoriously difficult to do, as much of the literature on teacher development suggests. It is highly unlikely that teaching could be improved significantly simply by placing a camera in the room. On the other hand, teachers will obviously try to do an especially good job, and may even do some extra preparation, for a lesson that is to be videotaped. We may, therefore, see a somewhat idealized version of what the teacher normally does in the classroom.

Minimizing Bias Due to Observer Effects

We have identified three techniques for minimizing bias due to videotaping. First, instructions must be standardized. Teachers generally do not want to bias the results of a study, but may inadvertently do so in an effort to help researchers. It is important, therefore, to clearly communicate the goal of the research to the teacher in carefully written, standard instructions. The teacher, when properly informed, becomes an important ally in the effort to get unbiased results. Teachers need to be told that the goal is to videotape a typical lesson, whatever they would have been doing had the videographer not shown up. Teachers can also be explicitly asked to prepare for the target lesson just as they would for a typical lesson.

A second technique is to assess the degree to which bias has occurred. After the videotaping, teachers can be asked to fill out a questionnaire in which they rate, for example, the typicality of what we see on the videotape, and describe in writing any aspect of the lesson they feel was not typical. We also can ask teachers whether the lesson in the videotape was a stand-alone lesson or part of a sequence of lessons, and to describe what they did yesterday and what they plan to do in tomorrow's lesson. Lessons that are stand-alone and that have little relation to the lessons on adjoining days may be special lessons constructed for the purpose of the videotaping. In the work we have done, however, this is rarely the case.

Finally, we must use common sense in deciding the kinds of indicators that may be susceptible to bias, and take this into account in interpreting the results of a study. It seems likely, for example, that students will try to be on their best behavior with a videographer present, and so we may not get a valid measure from video of the frequency with which teachers must discipline students. On the other hand, it seems unlikely that teachers will ask different kinds of questions while being videotaped than they would ask when the camera is not present.

Sampling and Validity

Observer effects are not the only threat to validity of video survey data. Sampling—of schools, teachers, class periods, lesson topics, and parts of the school year—is also a major concern.

One key issue is the number of times any given teacher in the sample should be videotaped. This obviously will depend on the level of analysis to be used. If we need a valid and reliable picture of individual teachers then we must tape the teacher multiple times, as teachers vary from day to day in the kind of lesson they teach as well as in their success in implementing the lesson. If we want a school-level picture, or a national-level picture, then we obviously can tape each teacher fewer times, provided we resist the temptation to view the resulting data as indicating anything reliable about the individual teacher.

On the other hand, taping each teacher once limits the kinds of generalizations we can make about instruction. Teaching involves more than constructing and implementing lessons. It also involves weaving together multiple lessons into units that stretch out over days and weeks. If multiple teachers are taped once, it will be difficult to code the dynamics of teaching over the course of a unit. Inferences about these dynamics cannot necessarily be made, even at the aggregate level, based on one-time observations.

Another sampling issue concerns representativeness of the sample across the school year. This is especially important in cross-national surveys where centralized curricula can lead to high correlations of particular topics with particular parts of the year. Although at first it may seem desirable to sample particular topics in the curriculum in order to make comparisons more valid, in practice this is virtually impossible. Especially across cultures, teachers may define topics so differently that the resulting samples become less rather than more comparable. Randomization appears to be the most practical approach to ensuring the comparability of samples.

Confidentiality

Unlike traditional data sets, much of the contents of video data will still be unanalyzed by the time a public-use data set is constructed. Yet, the fact that images of teachers and students appear on the tapes makes it even more difficult than usual to protect the confidentiality of study participants. An important issue, therefore, concerns how procedures can be established to allow continued access to video data by researchers interested in secondary analysis.

One option is to disguise the participants by blurring their faces on the video. This can be accomplished with modern-day digital video editing tools, but it is expensive at present to do this for an entire data set. A more practical approach is to define special access procedures that will make is possible protect the confidentiality of participants while still making the videos available as part of a restricted-use data set. (One such set of procedures is outlined below.)

Expense/Logistics

Video surveys can be far more expensive than traditional surveys. In fact, the future viability of such studies will depend on our ability to manage the considerable expense and logistical challenges posed by such studies.

Contrary to traditional surveys, which require intensive and thorough preparation up front, the most expensive and daunting part of video surveys is in the data management and analysis phase. Whereas information entered on questionnaires can easily be transformed into computer readable format, such is not the case for video images. Thus, it is necessary to find a means to index the contents of the hundreds of hours of tape that can be collected in a video survey. Otherwise, the labor involved in analyzing the tapes grows enormously.

Once data are indexed, there is still the problem of coding. Coding of videotapes is renowned as highly labor intensive. But there are strategies available for bringing the task under control. One approach to this task will be elaborated below.

TIMSS VIDEOTAPE CLASSROOM STUDY: SCALING UP TO VIDEO SURVEYS

Having discussed both the opportunities and the challenges offered by video surveys, we now turn to briefly describe an example of such a survey that is currently underway. This study, which is part of the Third International Mathematics and Science Study (TIMSS), represents an unprecedented attempt to use video in a national-level survey research context. Focused on 8th-grade mathematics, the study compares the teaching practices of German, Japanese, and American teachers. Data collection is complete; we are now coding the data. All of the issues described above have been encountered in the conduct of this study. Our experiences in addressing these issues will hopefully be instructive as we contemplate future video surveys.

Introduction to the Study

Background and Objectives

TIMSS is the third in a series of international studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement. The first two of these studies (Husen 1967; McKnight et al. 1987) established large cross-national differences in achievement, and provided some information on contextual factors, such as curriculum, that could be related to the achievement differences.

Perhaps because students from the United States did relatively poorly in the first two studies, the U.S. sponsors of TIMSS (primarily NCES) have placed a high priority on improving the quantity and quality of contextual information to be collected in TIMSS. Predicting that the performance of U.S. students would continue to be low relative to other industrialized countries, the U.S. Department of Education has tried to ensure that the results of TIMSS bear not only on the achievement of students but also on the processes that lead to achievement. The goal is to make TIMSS more useful to policymakers than either of the first two IEA studies have been.

In accordance with this goal, NCES has funded two studies to complement the main TIMSS data. Both of these studies focus on three countries: Germany, Japan, and the United States. The first involves comparative case studies of various aspects of the educational systems of each country. The second is the Videotape Classroom Study.

The goal of the Videotape Classroom Study is to provide a rich source of information on how 8th-grade mathematics is taught in Germany, Japan, and the United States. This is the first large-scale study to collect videotaped records of classroom instruction in the mathematics classrooms of different countries. The study has four main objectives:

1) To develop objective observational measures of classroom instruction that will serve as valid quantitative indicators, at a national level, of teaching practices in three countries;
2) To complement information about classroom instructional methods collected by the TIMSS background questionnaires with information gained from actual classroom observations in order to obtain a richer description of classroom teaching practices in Japan, Germany, and the United States;
3) To compare actual mathematics teaching methods in the United States and other countries with those recommended in current reform documents and with teachers' perceptions of those recommendations; and
4) To assess the feasibility of applying videotape methodology in future wider scale national and international surveys of classroom instructional practices.

Design of the Study

National probability samples of 8th-grade mathematics classes from Germany, Japan, and the United States are participating in the study. The samples are random subsamples of the TIMSS main study sample, which is selected according to the TIMSS sampling plan. The plan was to sample 100 classrooms from Germany and the United States, and 50 from Japan. The final sample consists of 100 classrooms from Germany, 81 from the United States, and 50 from Japan.

The video study includes two major sources of data: videotapes and questionnaires. In addition, supplementary materials helpful in understanding the lesson, such as examples of textbook pages and worksheets, were collected. Each classroom was videotaped once on a date convenient for the teacher. One complete lesson—as defined by the teacher—was videotaped in each classroom. One videographer was employed in each country. In Germany and the United States videotaping was carried out over a 7-month period, and in Japan, over a 4-month period. Teachers were told that we wanted to tape a "typical" lesson and, thus, that they should do no special preparation on the day of taping. After the taping, each teacher was given a questionnaire and an envelope in which to return it. The purpose of the questionnaire was to assess how typical the lesson was according to the teacher, and to gather contextual information important for understanding the contents of the videotape. Both taping procedures and questionnaire contents are described in more detail below.

The LAVA Software System

To facilitate the processing of such large quantities of video data, we decided to digitize all of the video and supplementary materials, which allowed them to be stored, accessed, and analyzed by computer. Each lesson videotape was digitized, compressed, and stored on CD-ROM disks, one lesson per disk. We then designed and built a multimedia database software application that would enable us to organize, transcribe, code, and analyze the digital video. This interactive video analysis system, which we have called LAVA (for LA Video Analysis), represents a major advance in technology available to aid in the implementation of video surveys. For this reason, the system will be described in some detail along with the description of each part of the study.

Digital video offers several advantages over videotape for use in video surveys. First, the resulting files are far more durable and long-lasting than videotape. CD-ROM disks are assumed to last for 100 years, as opposed to a much shorter lifespan for videotape. Digital video files also can be copied without any loss in quality, which again is not true for videotapes. And, digital files will not wear out or degrade with repeated playing and replaying of parts of the video. Digital video also enables random, instantaneous access to any location on the video, a feature that makes possible far more sophisticated analyses than are possible with videotape. For example, when coding a category of behavior, it is possible to quickly review the actual video segments that have been marked for that category. This rapid retrieval and viewing of coded segments makes it possible to notice inconsistencies in coding, or to discover new patterns of behavior, that would not be possible without such access.

The LAVA software system consists of several modes. Transcribe mode is used for transcribing the videotapes. Code mode allows users to define categories and code them across a large number of videos. Analyze mode is used to search the database and retrieve video segments on the basis of transcript or codes, and to produce spreadsheet outputs of data that can be imported into standard statistical analysis programs. These modes will be described in more detail later.

Instructions and Questionnaire

As pointed out earlier, both instructions to the teacher and the questionnaire that accompanies the videotaping are means of minimizing the potential bias of observer effects. Designing each of these was given careful consideration in the TIMSS video study.

Instructions

It is not feasible to show up unannounced to videotape classroom lessons. Because teachers know when the taping is to take place, they undoubtedly prepare for it in some way. How they prepare probably will have an impact on the kind of instruction we see. Teachers may try to teach like they think we want them to teach; they almost certainly will try to do what they believe is a good job.

In order to cut down somewhat on the variability in preparation methods across teachers, we gave teachers in each country a common set of instructions for how we wanted them to prepare. Teachers were told the following:

> Our goal is to see what typically happens in American mathematics classrooms, so we really want to see exactly what you would have done had we not been videotaping. Although you will be contacted ahead of time, and you will know the exact date and time that your classroom will be videotaped, we ask that you *not* make any special preparations for this class. So please, do not make special materials, or plan special lessons, that would not typify what normally occurs in your classroom. Also, please do not prepare your students in any special way for this class. Do not, for example, practice the lesson ahead of time with your students.

Questionnaire

The purpose of the teacher questionnaire was to elicit information that would help us in the analysis and interpretation of the videotapes. Items for the questionnaire were generated by project personnel in consultation with persons working on the main TIMSS questionnaire, questionnaire design specialists from Westat, mathematics educators, and classroom teachers. Questions were edited and selected to yield a questionnaire that would take approximately 20–30 minutes for teachers to complete.

The questionnaire was translated into German and Japanese, translated back into English, and then pilot-tested on teachers participating in the field test. The responses from the field test were discussed by German, Japanese, and American collaborators, and based on these discussions the questionnaire was revised.

The final translation of the questionnaire was painstakingly reviewed, question by question, by a group of German, Japanese, and American researchers, each of whom was fluent in two of the three languages. Questions that were judged too difficult to translate accurately were dropped from the questionnaire.

The resulting questionnaire consists of 3 parts with a total of 28 questions. In Part A, we ask questions about the lesson that was videotaped, and about how the class was constituted and who the students were. In Part B, we ask the teachers to compare what happened in the videotaped lesson with what would typically transpire in their classroom. In Part C, we ask teachers to describe what they know about current ideas on mathematics teaching and learning, and ask them to evaluate their own teaching in the videotape in light of these current ideas.

The information collected in the questionnaire will serve three purposes. First, information from the questionnaire will help us assess the quality and comparability of our samples across the three countries. Although teachers will be instructed not to prepare in any special way for the videotaping, we cannot take it for granted that what we see on the videotape is typical of what normally happens in a given classroom. Teachers thus will be asked to directly rate the typicality of the videotaped lesson, and these ratings will be compared across countries. Similarly, we will assess the comparability of the samples across the three countries along several important dimensions. For example, whether a lesson deals with new material or review might be expected to influence the kind of teaching technique used. Knowing the percentage of lessons in each country that are new versus review will help us to judge the comparability of the samples.

A second purpose for the questionnaire is to provide coders with information that will help them interpret what they see on the videotapes. For example, it is often necessary to know the teacher's goal for a lesson in order to make sense of the activities that constitute the lesson, and so we ask the teacher to say what her or his goal for the lesson was. Similarly, to interpret the meaning a specific question has for students it is often helpful to know whether the question probes new material or reviews previously learned information. Again, teachers are asked to categorize the content of the lesson in this way on the questionnaire.

Third, the questionnaire responses will, in some cases, enter directly into the analyses—statistical and qualitative—of the videotapes. This will occur in several ways. First, questionnaire responses will enter into correlational analyses within each country to help us relate contextual factors to variations in classroom instruction. For example, we can investigate the degree to which instructional techniques vary according to the ability level of students in the class. Second, we can use questionnaire responses to identify sampling biases that may affect our results. For example, if lessons that deal with new material (as opposed to review material) are sampled more in one country than another, this information could be used as a covariate to correct for the bias in sampling. Third, by asking teachers to comment on the lesson that was videotaped, we can learn more about how teachers interpret the language of reform in mathematics education. For example, if a teacher tells us that his or her lesson was focused on problem solving, we can look at the video to see what the teacher meant by the term "problem solving."

Filming in Classrooms

Before we could collect our first videotape, we had to accomplish a number of tasks. We had to 1) develop procedures for videotaping in classrooms that could be applied in comparable ways across three different cultures; 2) develop and implement methods for training videographers to use these procedures in a consistent way; and 3) evaluate the success of our training by comparing camera use across our three videographers. The following will describe how we accomplished each task.

Establishing Comparable Procedures

The success of any video survey will hinge on the quality and comparability of the tapes collected. What we see on video is not only dependent on what transpires in the classroom but also on the way the camera is used. If our aim is to compare certain aspects of instruction, then we must make sure that these aspects are clearly captured on all the tapes. In addition, we want to make sure that we are comparing classroom instruction, not camera habits. There are many decisions that must be made by the camera operator; if these are not made in a standardized manner, then the resulting tapes will not be comparable across classrooms or countries.

We developed procedures for camera use in collaboration with Scott Rankin, an experienced videographer who had worked with us on previous projects and who was therefore familiar with the challenges of documenting classroom instruction. Our goal was to develop a set of general principles and rules of thumb that would be easy for our videographers to learn, yet comprehensive enough to apply in any classroom situation. Of course, there are many rules and principles one could come up with depending on the goals of any particular survey. Reviewing ours, however, will at least serve to highlight the kinds of issues that must be considered when developing procedures for camera use. They might also be applicable to other studies.[1]

One camera was used, which of course limits the amount of information that can be collected. This constraint was imposed by NCES as a cost-saving measure, though it also makes the process of coding and analysis simpler than it would be with two cameras. The procedures for camera use presented would need to be altered if two cameras were used.

Basic Principles for Documenting Classroom Lessons

Because we wanted to see each lesson in its entirety, all videotaping was done in real time: the camera was turned on at the beginning of the class, and not turned off until the lesson was over. This means that we can study the durations of classroom activities by measuring their length on the videotape. Obviously, this would not be possible if there were any gaps in the recording.

Classrooms are complex environments where much is going on at any given time; it is impossible to document everything, particularly when only one camera is used. We decided on two principles to guide videographers in their choices of where to point the camera. These principles yield a comprehensive view of the lesson being taped.

Principle #1: Document the perspective of an ideal student. Assume the perspective of an ideal student in the class, then point the camera toward that which should be the focus of the ideal student at any given time. An ideal student is one who is always attentive to the lesson at hand, and always occupied with the learning tasks assigned by the teacher. An ideal student will attend to individual work when assigned to work alone, will attend to the teacher when he or she addresses the class, and will attend to peers when they ask questions or present their work or ideas to the whole class. In other words, we chose to point the camera so as to capture the experience of a student who is paying attention to the lesson as it unfolds. In cases where different students in the same class are engaged in different activities, the ideal student is assumed to be doing whatever the majority of students are doing.

Principle #2: Document the teacher. Regardless of what the ideal student is doing, be certain to capture everything that the teacher is doing to instruct the class. Usually the two principles are in agreement: whenever the ideal student is attending to the teacher, both principles would involve having the camera pointed at the teacher. However there are times when the two principles are in conflict. Take, for example, a case where the majority of students are doing seatwork while the teacher is working privately with two students at the board. The ideal student would be focused on his or her work, not on the teacher. In situations like this one, the videographer must go beyond these two basic principles in order to determine where to point the camera.

The Exceptions: Three Difficult Situations

We have identified three common situations where the principles alone cannot guide choices about what to capture on the videotape. These situations are 1) when the ideal student would be focused on something other than the teacher, 2) when two speakers who are having a conversation will not fit in a single shot, and 3) when a speaker and an object being discussed will not fit in a single shot. We have developed a set of guidelines so that videographers will chose similar (i.e., comparable) shots when faced with each of these situations, and so that these shots will contain a maximum amount of useful information. The rest of this section presents a more detailed discussion of these situations and how to film them.

Situation #1: When the ideal student is not watching the teacher. As already mentioned, there are times when the ideal student should be attending to something other than the teacher. This most often occurs when students are given a task to work on individually or in small groups. Teachers can use this time in different ways. Sometimes they will walk around the class and monitor students' work. This is ideal from the videographers' point of view because by following the teacher with the camera one can also get a sense of what students are doing. In some instances, however, a problem arises because the teacher does not circulate around the class, but rather stays at the board or his or her desk. In such cases, the camera would need to be pointed in two different directions (toward the teacher and toward the students) in order to capture both the teacher and the focus of the ideal student.

Videographers were instructed to handle such situations by alternating between these two points of view. They were told to slowly do a sweep of the classroom by panning away from the teacher and then panning back to the teacher so as to document what the students are doing. After this sweep, they were told to focus on the teacher unless the nature of the students' activity changes in any significant way (e.g., new materials are introduced or they break into groups). If the students' activity were to change, videographers were instructed to carry out another sweep of the students, and then return to the teacher.

Situation #2: When two speakers will not fit in a single shot. A second difficult situation occurs when the teacher is conversing with a student (or a student is conversing with another student) and the two speakers are far enough apart so that they do not fit in a single camera shot. This often occurs when a teacher calls on a student seated in the back of the room, and then proceeds to converse with the student.

In this case, videographers were instructed to move the shot from speaker to speaker as they take turns talking. An exception to this rule occurs when one of the speaker's turns is so brief that there is no time to shift the camera before the turn is over. In this case, the camera should be kept on the person doing the most talking.

Situation #3: When the speaker and the object being discussed will not fit in a single shot. Another difficult situation occurs when a speaker and an object he or she is discussing will not both fit into a single camera shot. This happens frequently, for example, when someone is talking about things written on the chalkboard or about concrete representations of a mathematical situation or concept.

In this kind of situation, videographers were told to document the object for long enough to provide the visual information needed to make sense of the talk, then to keep the shot on the speaker. For example, if the teacher is talking about a problem on the blackboard, the videographer should first tape the problem, then move to the teacher.

There is one important exception to this rule. Sometimes it is not sufficient to briefly see the object and then move to the speaker because the talk will make no sense unless one is seeing the object as it is being talked about. For example, if the speaker is pointing to specific features of the object as he or she talks, and if the direction of the points must be seen in order to understand the talk, then the rule is that the camera must stay on the object so that the talk can be understood.

How Close to Frame the Shot

Aside from making sure that videographers point their cameras at comparable things, we also wanted to make sure that their shots are framed in comparable ways. An extreme close-up of the teacher talking would provide a very different sense of the action taking place than a wide shot where the teacher is seen in the context of the classroom.

We decided that in general we wanted the widest shot possible, a shot professional videographers call the "Master of Scene" (MOS) or, more simply, the "master shot." From an aesthetic point of view closer shots often look better. However, the MOS provides more contextual information and thus was judged more appropriate for our purposes. The master shot also is less prone to bias because it does not artificially focus the viewer in on whatever aspect of the lesson the videographer judged to be most interesting.

Sometimes, however, there is crucial information that cannot be captured in a master shot. Common examples include objects being discussed during the lesson, or things written on the blackboard. In such instances, the camera should zoom in close enough to capture this information. In other words, although our preferred view of the classroom is the MOS, a closer shot must be used when it is needed to understand what is going on. Videographers were told to hold close shots long enough to enable a viewer to read or form a mental image of the information.

Moving from Shot to Shot

Finally, having devised guidelines for what to include in the shot, we also needed some rules for how to move from shot to shot. This, too, must be done in a standardized way if the tapes are to be fully comparable.

The guidelines we gave to the videographers were based on principles of good camera work. We taught them how to compose shots and execute camera movements in ways that follow basic cinematographic conventions and fundamentals of good composition. Aside from wanting them to follow the same conventions, we wanted them to carry out good camera work. Bad camera work calls attention to itself and distracts the viewer from the contents of the tape.

Training Videographers

In order to make sure that the rules were applied correctly and reliably, we had to work intensively with the videographers. Each videographer participated in two training sessions, both of which were conducted by our professional videographer. The first training session lasted 9 days in the spring of 1994, after which each videographer was sent out to collect ten practice tapes for a field test. The second training session lasted 5 days and was held in the early fall of 1994. Following this second training session, videographers were given a test, and then sent off to collect the data.

We designed the training sessions with two goals in mind: First, we wanted to teach the videographers our camera use rules to the point that they could follow them second nature. In an actual taping situation, videographers would have to make rapid decisions about where to point the camera without time for reflection. Second, we wanted the videographers to learn and practice the fundamental skills of camera use. These skills include, for example, changing from one camera angle to another quickly without losing a focused image, tracking moving objects without having the object leave the shot, and moving rapidly back and forth from close-ups to master shots, while ending up centered on the shot that needs to be captured.

The first training session was devoted to five activities: Learning to use the equipment, practicing basic principles of good camera work, presentation and discussion of the standardized rules for taping classrooms, practice taping in mock classrooms, and practice taping in real classrooms. Activities in the second training session included reviewing and discussing the rules, critiquing practice tapes, and more practice taping in mock classrooms. A monitor hooked to the camera during the training sessions allowed videographers to rotate between practicing with the camera and watching/critiquing their peers in collaboration with the instructor.

The following is a helpful hint for others contemplating this kind of work. One has two alternatives in deciding who to hire and train as a video survey videographer: one can hire scientists (i.e., educational researchers) and train them to take good pictures, or one can hire artists (i.e., photographers) and teach them the importance of following standardized rules for camera use. In my experience, the latter is far easier, and the pictures are much more aesthetically pleasing.

Evaluating the Comparability of Camera Use

At the end of the second training session, we gave each videographer a test to measure and document how well they had internalized all they had been taught. A 7-minute mock lesson was created that covered many of the situations videographers needed to know how to handle. The lesson was taught three times, each one identical to the others, and was taped each time by one of the three videographers. The resulting tapes were analyzed and evaluated to make sure that our videographers would shoot lessons in a standardized manner.

To evaluate the videographers' performance on the test, we first produced a description of how the test lesson should have been videotaped. We listed the 22 events that took place in the lesson, and then determined how each event should be taped given the procedures we had developed.

Once we had a description of how the test lesson should have been taped, we evaluated each videographer's performance against this ideal. We used a three-point scale to score how well they taped each of the 22 lesson events. The videographers were given a score of zero if they broke any of the rules that they needed to take into account. For example, if they did not zoom in to capture information that they were supposed to capture, or if they pointed the camera at the wrong thing, they would be given a score of zero. They were given a score of one if they showed an understanding of the rule they needed to carry out but did not apply it in a timely fashion. For example, if they needed to zoom in and capture what the teacher was pointing to but reacted too slowly and missed this information, or if they let the teacher walk around the class for a while before they decided to follow her or him, they would receive a score of one. They were given a score of two if they applied the rules exactly as we had predicted they should.

The scores obtained were all in a similar range and also were relatively high. The German videographer received a score of 35 out of a possible total of 44. The Japanese videographer received a score of 36, and the American videographer a score of 43. In addition, of the 66 events scored for the three videographers, only 4 were rated a zero (which means that a rule was actually broken only 4 times). Two of these zeroes were obtained by the German videographer, and two by the Japanese videographer. This means that no videographer ever showed more than two rule breaches for the entire test.

Performance on the test was also used to evaluate the quality of each videographer's camera work. First we generated a list of possible flaws that a videographer might produce. Our list included the following flaws:

- Cropping shots too tightly (e.g., cutting off part of someone's head).
- Cropping shots too wide (e.g., too much head room).
- Zooming in/out and then having to reframe the shot.
- Zooming in/out and then having to refocus the shot.
- Panning while zoomed in tightly.
- Jerky or awkward camera movement during zooms or pans.
- Losing from the frame any object that is being tracked.
- Unnecessary camera movement.
- Bad coordination between zooms and pans.
- Very unbalanced composition.

We used this list to score each videographer's performance on a four-point scale for each of the 22 events in the test lesson. Videographers were given a score of three on an event if we could find no flaw in their camera work. They received a score of two if one flaw could be found, a score of one if two flaws could be found, and a score of zero if at least three flaws could be found.

All videographers obtained scores that were within a similar range and judged to be satisfactorily high. The Japanese videographer received a score of 51 out of a possible total of 66. The German videographer received a score of 52, and the American videographer a score of 60. Both evaluations of the test confirmed our informal impression that camera standardization had been reached by the end of the training.

Videographers were in the field for a prolonged period of time. We worried, therefore, that they might slowly forget what they were taught or develop bad habits. In order to make sure that they continued using the camera correctly, every 10th tape that came in from the field was evaluated using a scoring system similar to the one described above. Videographers were given feedback about how they were doing. In particular, they were immediately informed if they had, in any way, drifted away from the standards we knew they were able to follow.

Gaining Cooperation from Teachers

We were concerned at the outset of the study that we would have difficulty finding teachers who were willing to be videotaped. Anticipating such difficulty, we decided to pay teachers for their participation. However, our fears may have been unfounded. In fact, getting schools to participate in the main TIMSS study proved to be more difficult than getting them to participate in the video study. I believe this is because the actual demands imposed by videotaping are minimal compared to those imposed by testing of students. As video surveys become more commonplace, it may prove easier and easier to secure cooperation from teachers, so long as videotaping is not tied to accountability for individual teachers.

Some Notes on Equipment

The quality of the data depends to a great extent on the quality of the equipment used in collecting the data. Thus, we wanted to use high-quality cameras that would produce excellent images, and high-quality microphones that would enable us to hear most of what goes on in the classroom.

The camera we selected was a Sony EVW-300 three-chip professional Hi-8 camcorder. Each camera was mounted on a Bogen fluid-head tripod. (Tripods that are not fluid head will produce jerky camera movements.) A small LCD monitor was mounted on the camera to help operators view what they were taping. Sound was collected using two microphones, one a radio microphone worn by the teacher, the second a shotgun zoom microphone mounted on the camera. Good audio is both difficult to achieve in classrooms, and extremely important for analyzing the contents of the tapes. Thus, it is best to purchase the highest quality microphones available.

Constructing a Multimedia Database

As the tapes and supplementary materials are collected, they are mailed to our project headquarters at UCLA. The tapes are then processed as follows: Videotapes and supplemental images are digitized, compressed, and stored on CD-ROM. Using software we have developed for this study, videotapes are transcribed, translated into English, and marked with time codes so that transcripts and video can be linked in a multimedia database. In the following sections we will describe these procedures in more detail.

Digitizing, Compression, and Storage on CD-ROM

The first step in constructing the multimedia database is to store the videotapes and supplementary materials in digital form on CD-ROM disks.

Because video contains so much information, it has until recently not been feasible to store large quantities of video in digital form. The breakthrough that makes such storage possible has been in the development of algorithms for compressing digital video so that it can be stored in smaller and smaller spaces. The algorithm we are using in the current project is called MPEG-1, an algorithm endorsed by the Motion Picture Engineers Group, that is fast becoming the industry standard. MPEG compression makes it possible to store 74 minutes of video and audio on a single CD-ROM disk.

Once we receive our videotapes, we digitize the tapes and compress them into an MPEG file on a large hard disk. Text pages, worksheets, and other supplementary materials collected by the videographers are digitized on a flatbed scanner and stored in PICT format on the same hard disk drive as the accompanying videotape.

Once the MPEG file and accompanying PICT files for each lesson are stored on the hard disk drive, the files are burned onto a CD-ROM.

Software and Hardware for Accessing Digital Video

Once the video is stored on CD-ROM disks, it can be accessed by the database software we have developed for this project. Users of the software work at a computer workstation consisting of the following:

- Apple Macintosh Power PC 8100AV computer with built-in CD-ROM drive;
- Apple 17-inch Multi-Scan monitor;
- Hardware card in computer for real-time decoding MPEG files (manufactured by Wired, Inc.); and
- Headphones.

Workstations are networked together in a client/server system. The server consists of a Macintosh Power PC 8100 computer. Although video is stored locally on CD-ROM at each workstation, all transcription/translation and time codes that link the transcription to the video are stored on a central server. This makes it possible for many transcribers and coders to work simultaneously on a single, integrated database. It also means that later, in the analysis phase, we will be able to apply sophisticated search procedures to the entire database at once, without having to change CDs. Only if we need to view the video itself will it be necessary to locate and load the actual CD.

We have so far implemented three modules in the software: transcribe, code, and analyze.

The transcribe module enables transcribers/translators to:

- View the video and control playback through a window on their computer screen;
- Type the transcription/translation into another window on the screen; and
- View the video, once transcribed, with subtitles in real time.

The transcriber sees two major windows on the computer screen: one displays video, the other displays the transcript. Under the video window is a rectangular area used for displaying transcript records as subtitles in real time, and various buttons for controlling the video. Various controls allow the transcriber to:

- Set up and easily modify a continuous loop so they can watch the same segment of video over and over while they transcribe/translate the speech;
- Move the loop forward to continue transcribing the next segment of video;
- Stamp time codes to mark the beginning of each utterance;
- Enter new records into the transcription database;
- Merge records together and break records apart;
- Move instantly to the point in the video that corresponds to the highlighted transcript record;
- Move instantly to the point in the transcript that is closest in time to the point where the video currently is; and
- Turn synchronized subtitles on and off while viewing the video.

Transcription/Translation of Lessons

Our goal is to have transcripts that reflect, as accurately as possible, the words spoken by both the teachers and the students. It is not enough to summarize or paraphrase the talk, nor is it acceptable to transcribe the data in a way that reflects what the participants mean to say.

We have developed a protocol to make sure that all transcription/translations are carried out in a standardized manner. For example, transcribers are given rules about how to indicate speakers, how to break speech into turns, how to use punctuation in a standardized manner, and how to translate technical terms in a consistent way.

Each American lesson is transcribed in order to facilitate coding. Because some parts of the video are hard to hear, the transcript enables the coder to better understand what is happening in the lesson. It also is possible to code some aspects of instruction directly from the transcript, without viewing the video at all.

German and Japanese lessons are translated into English as they are being transcribed. The purpose of the translation is to aid in multilanguage searches of the database, and to make it possible for persons not fluent in German or Japanese to view and understand the lessons. All coding of the videotapes will be done by native speakers of the language being coded. Thus, coders will not rely on translations to make subtle judgments about the contents of the video.

Videotapes are transcribed and translated by teams of transcribers fluent in each of the three languages. Some members of the German and Japanese teams are native speakers of those languages, others are native speakers of English but fluent in German or Japanese. Each tape is transcribed/translated in two passes. One person will work on the first pass transcription/translation of a tape, and then a different person is assigned to review this work. A hard copy of the first pass transcription/translation is printed out, and the reviewer marks any points of disagreement on this copy. The two individuals then meet, discuss all the proposed revisions, and come to an agreement about what the final version should be. In cases where disagreements cannot be resolved, a third party is consulted.

The last step in the transcription/translation process is to time code the tapes, i.e., to mark the exact point at which each utterance begins.

Coding and Analysis

Instructional quality is a complex construct for which few standard indicators exist. Coding of classroom videotapes, therefore, is part of a cyclical process that involves refining the construct, developing indicators of the construct, validating the indicators, and then using the results to further refine the construct. The state of the art of this process is at a very rudimentary level: we have poor ways of describing classroom processes at present. Partly this is because classroom instruction is a highly complex system that is inherently difficult to describe. It is also true that we have devoted far less energy to this enterprise than to measuring the outcomes of instruction.

This section will provide a description of how we began to develop the coding system for the TIMSS video study, and how we are implementing the coding in our LAVA software program.

Deciding What to Code

In deciding what to code, we had to keep two goals in mind: first, we wanted to code aspects of instruction that relate to our developing construct of instructional quality; second, we wanted the codes we used to provide us with a valid picture of instruction in three different cultures. For the first goal, we sought ideas of what to code from the research literature on the teaching and learning of mathematics, and from reform documents—such as the *NCTM Professional Teaching Standards*—that make clear recommendations about how mathematics ought to be taught. We wanted to code both the structural aspects of instruction, i.e., those things that the teacher most likely planned ahead of time, and the on-line aspects of instruction, i.e., the processes that unfold as the lesson progresses.

The dimensions of instruction we judged most important included the following:

- *The nature of the work environment.* How many students in the class? Do they work in groups or individually? How are the desks arranged? Do they have access to books and other materials? Is the class interrupted frequently? Do the lessons stay on course, or do they meander into irrelevant talk?
- *The nature of the work that students are engaged in.* How much time is devoted to skills, problem solving, and deepening of conceptual understanding? How advanced is the curriculum? How coherent is the content across the lesson? What is the level of mathematics in which students are engaged?
- *The methods teachers use for engaging students in work.* How do teachers structure lessons? How do teachers set up for seatwork, and how do they evaluate the products of seatwork? What is the teacher's role during seatwork? What kinds of discourse do teachers engage in during classwork? What kinds of performance expectations do teachers convey to students about the nature of mathematics?

Our second goal was to accurately portray instruction in Germany, Japan, and the United States. Toward this end, we were concerned that our description of classrooms in other countries make sense from within those cultures, and not just from the American point of view. One of the major opportunities of this study, after all, is that we may discover approaches to mathematics teaching in other cultures that we would not discover in our culture alone. We wanted to be sure that if different cultural scripts underlie instruction in each country, we would have a way to discover these scripts.

For this reason, we also sought coding ideas from the tapes themselves. In a field test, we collected nine tapes from each country. Collected in May 1994, we convened a team of six code developers—two from Germany, two from Japan, and two from the United States—to spend the summer watching and discussing the contents of the tapes in order to develop a deep understanding of how teachers construct and implement lessons in each country.

The process was a straightforward one: we would watch a tape, discuss it, and then watch another. As we worked our way through the tapes, we began to generate hypotheses about what the key cross-cultural differences might be. These hypotheses formed the basis of codes, i.e., objective procedures that could be used to quantitatively describe the videotapes. We also developed some hypotheses about general scripts that describe the overall process of a lesson, and devised ways to validate these scripts against the video data.

Developing Coding Procedures

Once the list of what to code has been created, we are ready to begin developing the specific procedures to be used in coding the tapes. First, field-test tapes are viewed by the coding development group, and a definition of the category to be coded is proposed. Then, code developers try to apply the definition to the field-test tapes from their country. Difficulties are brought back to the group, and definitions are revised and refined. This process is repeated until all members of the group are satisfied with the definitions and procedures, and agree with the coding of each instance.

Once codes are developed, coders are trained to implement the codes. Before coding begins, a formal reliability assessment is conducted to ensure independent agreement across coders at a level of at least 80 percent for each judgment. Reliability is assessed by comparing each coder's results with a standard produced by the coding development team.

Throughout this process we endeavor to be strategic. For example, just having collected 100 hours of video does not mean that all 100 hours must be analyzed. Depending on the frequency of what is being coded, it may be possible to time sample or event sample, and our computer software makes this easy to do. It is also important to divide coding tasks into passes through the data in order to lessen the load on coders. This increases reliability and speeds up coding.

Implementation of Codes Using the Software

The code module of our software enables coders to view synchronized video and transcript on their computer screen. On-screen controls allow them to move instantly to the point in the video that corresponds to the highlighted transcript record, or to the point in the transcript that is closest in time to the current frame of video.

Coders can work from video, transcript, or both, and they can mark the occurrence of events they are targeting in a given coding pass.

There are three types of events that can be coded:

1) In only—an event is marked by a single time point. Events would be coded this way when we do not care to measure their duration but just want to record their occurrence.

2) In and out—an event is marked with a beginning and end point on the videotape. Most of the events we code are of this type. For example, we code when periods of seatwork begin and end.

3) Exhaustive segmentation—a tape is segmented such that the end point of one segment serves as the beginning of the next, meaning that no part of the tape is not included in a segment. We use this type of event when coding classroom organization, for example. Coders are forced to categorize each part of the tape into one of the three categories of organization.

The software enables coders to code events from video by marking a beginning and ending, or beginning only, time code; or from transcript by marking the beginning and ending, or beginning only, points in the transcript. It also allows us to define new event types by searching Boolean combinations of other events and characteristics that have already been coded.

The software also allows the coder to characterize an event that has been coded. A button on the screen takes coders to the next event that has been coded, plays the event, and then presents the options for coding of characteristics. There are four types of characteristics that can be coded.

1) Numerical—an event is characterized by a numerical value on some dimension.
2) Mutually exclusive—an event is categorized into one of a mutually exclusive and exhaustive set of categories.
3) Check all that apply—an event is judged as belonging to one or more of a set of non-mutually exclusive categories.
4) Descriptive—a qualitative description is written and attached to a particular event.

Codes can be applied using one of four sampling schemes.

1) Play all—the coder can watch the entire lesson, marking codes whenever they are appropriate.
2) Play events—the coder can watch only events of a particular type, then characterize the events.
3) Sample events—the coder can be presented with a randomly chosen sample of events of a particular type.
4) Sample time—the coder can be presented with a randomly chosen sample of time segments, then mark whether or not specific events happened during each segment.

First-Pass Coding: The Lesson Tables

We have found that it is useful to have an intermediate representation of each lesson that can serve to guide coders as they try to comprehend a lesson, and that can be coded itself. For this purpose, our first step in coding the lessons is to construct a table that maps out the lesson along the following dimensions:

- Organization of class—each videotape will be divided into three segments: pre-lesson activities, lesson, and post-lesson activities. The lesson needs to be defined in this way because the lesson will be the basic unit of analysis in the study.
- Organization of interaction—the lesson is divided into periods of classwork and periods of seatwork.
- Activity segments within classwork—each classwork segment will be further divided, exhaustively, into activity segments according to changes in pedagogical function. We

have identified seven different kinds of activity segments: introduction, instructing, setting up seatwork, sharing seatwork product, correcting homework, test-taking, and conclusion.

- Activity segments within seatwork—we have distinguished three types of activities during seatwork: working on tasks and situations, correcting homework, and correcting seatwork. In addition, we have added two categories to characterize the kinds of simultaneous activities we have seen thus far: working and correcting homework, and working and correcting seatwork.
- Mathematical content of the lesson—the mathematical content of the lesson is divided into units. The content of each unit will be written down concretely/qualitatively, and then categorized into one of four types: situation, task, information, and solution method.

We are using these first-pass tables for two purposes. First, they can be used by subsequent coders to get oriented to the contents of the videotapes. Often it takes a great deal of time for coders to figure out what is happening in a lesson. The tables ease the way, providing an overview of the structure and content of each lesson.

A second purpose for the tables is that some codes can be coded from the tables without even going back to the videotapes. Examples of such codes include TIMSS content category, nature of tasks and situations, and changes in mathematical complexity over the course of the lesson.

Confidentiality and Sharing of Data

As pointed out above, there is a major issue concerning how to make video data available for secondary analysis while at the same time protecting the confidentiality of study participants. We have outlined one approach to accomplishing these goals as part of a proposal to establish the TIMSS video data as a restricted-use data set.

Our strategy for preserving the confidentiality of participants will be similar for both raw and restricted-use data sets. In general, we will separate the activity of coding the visual images (e.g., access to video pictures of teachers and students) from the activity of analyzing the results of the coding. Persons engaged in coding will have no access to any identifying information about teachers or students. They will know which country the teachers are from, but nothing else. Persons engaged in analysis, on the other hand, will work with data sets in which summary variables from the coding have been linked, via a teacher ID, to other information from TIMSS. But these analysts will not have access to video images.

This will be accomplished by constructing two independent data sets, one for the video data, the other for all other data. Separate ID numbers will be assigned to teachers in each data set. Information that can match IDs from one data set to the other will be held in a secure place, available only to senior personnel. A third, integrated data set will be constructed once we are ready to undertake integrated analyses. This integrated data set will not contain any visual images.

For the restricted-use data set, additional safeguards would be taken to make it practically impossible for researchers to link the two data sets with identifying information.

First, all specific identifying information would be deleted from the second data set; researchers would be provided with only a subset of variables that were available in the raw data set. For example, geographic region of the country would be deleted, as would size of school, age of teacher, and so on.

Second, we would exercise controls over the coding of video data that would prevent researchers from linking any specific image with any other data, although codes, of course, would be linked. We propose using the following procedures:

- Access to video data would be allowed only in specifically designated research rooms in which the full data set would be available. Researchers could view and code video data in this room or rooms, but would have no access to the second data set at all while they were coding video data. Researchers would not be allowed to remove any written materials from this room.
- After researchers complete their coding of the video images, project staff would construct aggregate data sets containing the results of the coding, remove all ID numbers, and then give the data back to researchers in an electronic spreadsheet format for analysis. Researchers who wanted additional TIMSS data integrated into their video coding spreadsheet would simply request that project staff put the additional variables into their spreadsheet. Again, all ID numbers would be deleted.

We believe that these safeguards would provide a high degree of confidentiality to participants while at the same time allowing researchers to access this valuable and unprecedented data set. Of course, if a researcher brought up an image and said "Oh, that's my sister-in-law," confidentiality would be undermined. But such an event is unlikely.

CONCLUSION

I began this paper by urging a new emphasis on developing and using observational indicators of classroom processes. I proposed video surveys as a promising approach to this task, but outlined some difficult issues in the implementation of video surveys. Finally, I showed how, in the TIMSS Videotape Classroom Study, we have successfully resolved these issues in the first large-scale video survey of classroom instruction.

Although I believe I have shown that video surveys are logistically possible, it is too early to see what the full benefits of such studies will be. The technology for assessing student outcomes has been developed over a long period of time. Research on classroom processes, in contrast, is still in its infancy. There is much work to be done before statistically acceptable, useful indicators are in hand. The task of developing such indicators, however, strikes me as one of the most important to be undertaken over the next decade. If we cannot make significant progress on the assessment of instructional processes, we will not have the basis on which to improve classroom instruction. Without this solid empirical foundation, efforts to reform instruction will continue to be grounded in ideological debates and pendulum swings.

NOTES

1. A more detailed account of these procedures can be found in the "TIMSS Videographers' Handbook," available by request from the author.

REFERENCES

Cronin, J.M. 1985. "Issues in National Educational Data Collection," NCES Redesign Project.

Hall, G., Jaeger, R. M., Kearney, C.P., and Wiley, D. E. 1985. *Alternatives for a National Data System on Elementary and Secondary Education.* Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Husen, T. 1967. *International Study of Achievement in Mathematics.* New York: Wiley.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kifer, E., Swafford, J.D., Travers, K.J., and Cooney, T.J. 1987. The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective. Champaign, IL: Stipes.

Peterson, P.L. 1985. "The Elementary/Secondary Redesign Project: Assessing the Condition of Education in the Next Decade," NCES Redesign Project.

Porter, A.C. January-February 1995. "The Uses and Misuses of Opportunity-to-Learn Standards," *Educational Researcher* 24 (1): 21–27.

Discussant Comments

KEVIN F. MILLER

NCES has supported developing a new technology that offers the promise to revolutionize our understanding of the processes that go on in classrooms and, in turn, to dramatically increase the impact of the research NCES supports. In this comment, I will 1) describe some of the consequences of this new technology; 2) discuss some aspects of human cognition that make it particularly important; and 3) argue that NCES could play a pivotal role in creating a new American Education Yearbook, including a video archive of educational processes in American schools.

Videosemantics: Making Sense Out of Classroom Processes

NCES collects data on teachers and classrooms as a method of describing the changing face of instruction in the United States and making it possible for researchers and policymakers to understand the instructional processes that account for changes in educational achievement. This is primarily done through surveys of teacher's beliefs, attitudes, and activities. As Stigler (1995) notes, there are fundamental problems in moving from these data to a real understanding of what goes on in classrooms. Self-reports of teaching practices may not produce accurate descriptions of actual classroom processes, because teachers may vary in how they interpret survey questions and may have limited and selective recollection of what transpires in their classrooms.

There is a more fundamental obstacle to going from surveys to prescriptions for improving instruction. In the same way that knowing the ingredients in a cake does not by itself enable you to bake one, knowing the characteristics of a good teacher does not in itself tell you how to become one.

What is needed to move from descriptions to prescriptions is a method of making the process of instruction explicit, and this is precisely what the video survey technology provides. The actual process of instruction can be made accessible to scientific study in a way that has been hitherto impractical. Observers could easily watch how 20 different teachers teach the same content, or how the same teacher responds to the questions of different students. Teachers in training could observe how skilled teachers respond to problems that come up in the course of instruction, and could watch themselves as they attempt to teach a lesson.

In his paper describing this technology, Stigler (1995) notes that the key to the revolutionary improvement in manufacturing quality engineered by W. Edwards Deming was the insight that improving quality requires one to focus on the processes of manufacturing rather than simply inspecting the products of those processes. NCES has made possible the development of technology that could lead to a similar revolution in education, by changing the focus from testing students and surveying teachers to actual measurement and description of instructional processes.

Taming the Power of the Anecdote

Video technology may also provide a solution to one of the most vexing problems facing educational researchers: the enormous difficulties that the consumers of research have in understanding statistical data. People are much more likely to be swayed by individual anecdotes than they are by carefully collected, representative data. A good demonstration of this problem was provided by Borgida and Nisbett (1977), who presented University of Michigan psychology undergraduates with evaluative information about upper level courses in their field. This was either presented by previously unknown confederates as representing their personal experience, or as the ratings of an entire group of students. Despite what these students must have learned about the effect of sample size on the reliability of observations, the individual reports had a significantly larger impact than did the statistical data on whether or not students planned to sign up for the recommended courses and avoid the non-recommended courses.

At its most extreme, the power of the anecdote that suggests the pictures appearing on the cover of an NAEP report may have more impact than the data contained inside. Statisticians may bemoan the power of anecdotal experience, yet it appears to be a fundamental aspect of human cognition. Video technology offers a potential solution here, providing a means for turning vignettes into data that can be presented systematically. Observers can code a corpus of classroom observations, producing quantitative descriptions of the data set. These quantitative descriptions can be coupled with presentations of examples of the kinds of processes observed. Because these observations are culled from a data set, it is possible to determine whether they are representative or exceptional, and it is possible for researchers with different interests to code the same data set in different ways. The melding of statistics and anecdotes that the video technology makes possible can be both powerful and methodologically responsible—powerful in the way that only direct experience can be, and responsible in that the statistical representativeness of these experiences can now be assessed.

Exploiting the Technology: A Yearbook of American Education

NCES has supported the developing of a revolutionary method for collecting educational data and making it accessible to researchers. It has an equally vital role to play in promulgating this technology and ensuring that it is used to understand the changing state of instructional processes in the United States. Imagine how valuable it would be if there were a systematic filmed record of teaching in the United States from earlier eras. Such a database would be a gold mine for researchers interested in all aspects of changes in the lives of children and the processes of education. A database of current instruction in the United States will be equally valuable for anyone who wishes to understand the changing face of schooling in America. NCES has experience in and sampling of the state of education in the United States. It should be within both its expertise and mission to develop a video yearbook of American education by collecting a representative sample of teaching in the United States. Such a database would be of interest to researchers and policymakers from a variety of fields. It would not only provide a vital record of the state of American education but also would be extremely useful in helping us to understand the classroom processes that result in effective instruction. Additionally, it would form a lasting legacy for future generations, who will use it to answer questions that we cannot now anticipate.

References

Borgida, E., and Nisbett, R. E. 1977. "The Differential Impact of Abstract vs. Concrete Information on Decisions." *Journal of Applied Social Psychology* 7: 258–271.

Stigler, J. W. 1995, November. "Large-Scale Video Surveys for the Study of Classroom Processes. Paper presented at the NCES conference "Future NCES Data Collection: Some Possible Directions." Washington, D.C.

8

Education for Work: Curriculum, Performance, and Labor Market Outcomes

Education and Work: Curriculum, Performance, and Job-Related Outcomes

---

Peter Cappelli

EXECUTIVE SUMMARY

Perhaps the most fundamental question within the topic of education and work is whether the two are in conflict. Are the requirements for success in the workplace in conflict with the goals of academic achievement? Putting this issue to rest would be an enormous contribution, but it requires data of the kind outlined below.

CHANGES IN THE WORKPLACE

Evidence suggests that skill requirements are clearly rising for many jobs, perhaps for the average, but not uniformly. The skills that are in increasing demand are often the kind of behavioral skills that have not typically been part of academic achievement assessments.

Declining attachment between employer and employee raises questions as to where workers will get skills. It puts the burden more on the education system, as we should expect people to go back and forth from school and work, repeating some of the school-to-work transition issues over a lifetime.

WHAT DOES WORK DEMAND OF EDUCATION?

We do not really know the answer to this question because most exercises simply ask employers whose requirements are always in flux. We need ways to validate estimates of the effects of education on actual job and organizational performance.

To do this, we need better data in three areas:

1) Identifying the knowledge, skills, and abilities (KSAs) produced by education, especially those traditionally excluded in academic achievement assessments;
2) Identifying the characteristics of schools and education that produce the desirable KSAs; and

3) Measures of performance in the workplace that go beyond wages in order to examine both the success of individuals and their organizations.

HOW DOES WORK AFFECT EDUCATION?

Longitudinal data are needed that go beyond the simple cross-sectional studies aimed at secondary school academic achievement, especially as we are increasingly concerned with lifelong learning. Again, we need better data in three areas:

1) Data and analyses relating work experience to postsecondary achievement, broadly defined;
2) The effects of work on a wide range of learning, including work-based skills, behavioral skills, and so on; and
3) More complete information on work experience, including the nature of tasks performed and the learning experiences at work.

SPECIFIC IMPLICATIONS FOR DATA COLLECTION

We need to link information about work experiences and education experiences in the same data sets, as well as more thorough measures of inputs and outcomes for both education and work in these data sets. The best approach is to leverage off of existing data sets as follows:

- Additional information on work experiences could be added to the NLS–72 and HS&B data sets;
- New longitudinal surveys are needed to collect more data on educational experiences and outcomes, especially for secondary school; and
- Data on employers represents the biggest challenge in order to understand how employer practices affect later education and how education and KSAs, in turn, affect organizational performance. More targeted surveys that match employers and employees might be the most cost-effective approach.

It has long been understood that education has an important influence on success in the workplace. More recently, many observers believe that this influence is becoming more important and that the benefits of education may well extend beyond the success of an individual worker to that of organizations and entire economies. With this visible change has come increased interest in exactly how education affects workplace performance. For example, does the subject matter and the pedagogy used affect workplace outcomes in addition to the credentials one attains, and are there innovations that could be made in the education system that would strengthen the relationship between education and performance?

The potential effects that work can have on education, on the other hand, have perhaps been less appreciated outside of the research community. The interest in work-based learning and the identification of skills that are best learned in the context of actual workplace experience are examples of the type of effects that work can have on education and learning.

The education research community has not always been especially interested in relationships with work. The understanding of the role education plays in labor market success, for example, was identified and championed by labor economists who were interested in understanding wages; the role that KSAs conveyed in education play in determining job performance was identified and researched by personnel psychologists whose goal was understanding effective employee selection; and the importance of work-based skills and learning has largely been advanced by studies of international competitiveness that emphasized the role that apprenticeships and other school-to-work programs play in raising national skill levels. The effects of work on traditional education have been perhaps more thoroughly examined by education researchers, although here the focus has often been either negative (linking student hours of employment to poor academic achievement) or highly focused on legislated programs such as review of vocational education programs mandated by the Perkins Act.

Perhaps the main reason for the relative lack of interest—and in some cases antipathy—in the workplace among the community of education practitioners and scholars was the sense that what mattered for workplace success was different, perhaps even antithetical, to the factors that shaped academic success. Scholars like Bowles and Gintis (1976) asserted that what employers wanted and perhaps needed from schools in terms of the characteristics of graduates/new hires was a kind of compliant behavior that was in conflict with the goals that educators held for their students. There are certainly arguments and evidence suggesting that they are not in conflict, but the view that they are remains deeply held in many circles.[1]

One very important consequence of this perception of conflict in goals has been continued support for an inward orientation toward evaluation in education: The "success" of an education establishment, for example, has almost uniformly been based on how well its students learn the material that educators have presented as assessed by the education community itself or, at the secondary school level, on how well graduates do in getting access to postsecondary education. Whether learning that material contributes in some important way to other life outcomes is rarely examined. Consider, for example, what the equivalent arrangement sounds like in a different situation like medicine. Procedures would be evaluated based on whether they did what the doctors wanted them to do and not necessarily whether it furthered the patient's health. "The treatment was a success but the patient died" is the aphorism used to parody such arrangements in medicine.

One of the first general priorities for NCES and the research community should be to address whether the goals of educators for students are in fact in conflict with the goals of workplace success. Specifically, whether achieving in school based on traditional measures is related to or in any way in conflict with achieving in the workplace. If the perception of conflicting interests can be put to rest, then at least some of the conflict between business and education may abate as well. This will also apply to at least some of the resistance to evaluation based in part on workplace outcomes in the education community, as well as the lack of real participation and commitment to education among the employer community. If, on the other hand, conflicts are identified between these goals, such information would provide important evidence for striking compromises or creating new arrangements for advancing success in both arenas.

Several factors are pushing the education community away from internal assessments and toward evaluations that are based more on external criteria. These efforts are widespread, and some are likely to be much more productive than others. On the negative side, they include pressures for "accountability" in the public sector, which have played out in postsecondary education as efforts to judge the efficiency of state education systems in crude cost-benefit terms, graduates per dollar. Attention to the workplace success of school leavers as part of the assessment of education is a potentially more useful development. It has been powerful for several reasons.

First, employers and policy observers have been vocal in their belief that the poor preparation of school leavers has contributed to problems inside organizations and in the economy as a whole (the extent to which they are right in thinking so is another matter). The reports making these arguments are so well known as to be almost household names. They include *A Nation At Risk* (1983), *Workforce 2000* (1985), and *America's Choice* (1989). The legislation that resulted in part from these arguments could institutionalize the interest in relating work to education, subject, of course, to continued funding from Congress. These include the School-to-Work Opportunities Act of 1994, with its efforts to develop infrastructure at the state level to bring school and work closer together, and the mandate of National Goals for Education Act of 1992 to develop national skill standards for jobs that can then be translated into curricula and credentials for participants.

Second, as the job market tightens, students and their parents will increasingly demand that schools—primarily postsecondary institutions—do a better job in preparing students for the workplace. Entry-level wages for college graduates have been falling rapidly in real terms—much more so than for the work force as a whole—while the proportion of college graduates who will find jobs requiring college skills is projected to decline.[2] The anecdotal reports from state university systems that as many as 25 percent of 4-year graduates return to community colleges for work-based classes before getting a job suggests something about the magnitude of the problems in preparing students for the workplace.

CHANGES IN THE WORKPLACE[3]

Behind the above pressures on education are profound changes in the workplace that will make very different demands on education systems and, more to the point, increase the importance of education suppliers to workers and the economy. The first of these is the change inside organizations as to how work is organized. Specifically, what new tasks are workers required to perform, and what different skills do those tasks demand from them?

Whether new models of work organization are in fact changing skills—and, if so, in what way—is a central question for advocates who believe that we need different kinds of data for research. Whether skill requirements are a more important issue now, where the kinds of skills that are important have changed, and whether these skills challenge existing data collection efforts are among the issues driven by workplace changes.

The place to begin that discussion is by asking what is happening, on average, to job requirements. Are skill demands really changing as much or in the manner that many advocates suggest, creating real shortages of workers with the education level necessary to fill jobs? The recent EQW/Census National Employer Survey (1995) found a majority of employers asserting that overall skill requirements have risen in their organizations for production or front-line jobs. But it may not be obvious how valid these responses are given the subjective nature of the responses where "skill" is not defined, for example, and can easily be confused with performance requirements. In other words, more may be demanded from employees, but what is really being increased is effort, not skill.

We used the EQW survey to examine what factors seem to differentiate those establishments reporting that skill requirements have risen for their front-line workers (Cappelli 1995). Those that have Total Quality Management Programs (TQM), more extensive teamwork arrangements and greater use of computers for both managers and non-managers, report that skill needs are increasing. These changes are consistent with the arguments that the shift toward "high-performance" workplaces is raising the skills needed in establishments that introduce those practices. As these practices become more widespread, these developments could have economy-level consequences. Establishments with more educated workers are also more likely to report that skill requirements are rising. This result is consistent with the arguments made by Bartel and Lichtenberg (1987) that more educated work forces have a comparative advantage in adopting innovations in technology and practices that might raise skill needs.

A different approach might be to look within establishments at the actual changes going on in the way work is organized. Consider, for example, the issue of autonomy, a key concept in participative work systems and an important factor in raising skill requirements. The argument is that as participative and decentralized work systems expand, employees have much greater autonomy in decision making and therefore need much greater skills to make the kind of decisions that their more highly trained supervisors had made for them in the past. But as Klein (1989) observes, just-in-time inventory systems that eliminate buffers of materials or intermediate products between work groups make those groups highly interdependent; changes in the production arrangements within any individual group can change its work pace, causing either shortages or pile-ups of material downstream. Because the overall flow of work across *all* teams in the assembly process must be absolutely consistent, the autonomy that any individual worker or team has to make changes in work organization is tightly constrained.

Further, as Adler (1993) discovered at the New United Motors (NUMMI) joint venture between Toyota and General Motors, the principle of continuous improvement requires that the performance of individual tasks be completely routinized so that the work teams can discover whether minute changes in tasks lead to an improvement in performance. In this sense, continuous improvement in work processes is like a laboratory experiment where everything is held constant except the one change being investigated. For employees, individual tasks appear to be every bit as rigidly defined as under scientific management. Individual workers in fact do not have the kind of autonomy that demands higher skill levels. The fact that the work teams themselves can influence the design of those tasks may make the system more palatable, however. In manufacturing, therefore, where most of the reform efforts have been concentrated, innovative production processes may not necessarily lead to work organization that makes dramatically different demands on production employees.

My study of changes in skill requirements used data obtained on 56,000 production workers over an 8-year period to examine whether skill requirements have changed. The results suggest significant upskilling for production jobs across the board as measured by changes in Hay points, the job evaluation metric used by Hay Associates to measure job requirements. Some of the upskilling seems due to the fact that tasks associated with quality control and housekeeping have been pushed onto all the remaining jobs (the decline of employment in quality and housekeeping jobs is consistent with this interpretation). That is, not only has each job experienced upskilling but also the overall distribution of production jobs has shifted away from less skilled and toward more skilled positions (Cappelli 1993).

"Lean production" techniques that have become popular in manufacturing (see below) essentially eliminate some jobs and push their tasks onto production workers. Some of those tasks, such as housekeeping, add little to the job. Other tasks, such as coordinating job design changes across teams, demand considerably higher skills, especially behavioral skills (communication, negotiation, and group dynamics skills). Adler (1993) notes that many of the tasks previously performed by industrial engineers, such as job analysis and redesign, are now being pushed down to the production teams.

It is also important to remember that while these skill requirements are rising, they start at a low base. Data from Hay Associates suggest that a typical management job, for example, has skill levels about twice those represented by production work. Given the low base, it is certainly possible that workers already have the skills to meet the increasing skill demands represented by these data. In other words, the fact that job requirements are rising does not necessarily mean that workers' existing skills are likely to be challenged.

Is There "Upskilling" Outside of Production?

By definition, the techniques of high-performance *production* systems are associated with production work, and not all of these techniques apply directly to other industries. The equivalent study to the one noted above using Hay data for clerical jobs finds no consistent pattern; some clerical occupations show increases in skill while others experienced decreases (Cappelli 1993).

One important attribute of the "lean-production" or "high-performance" work systems that do seem to raise skill requirements in manufacturing is the increased flexibility needed to handle variations in products. Situations that do not demand change—indeed may punish it—may not make great use of these techniques. There is relatively little use of high-performance production techniques in industries like transportation, distribution, or public utilities, perhaps because reliability and consistency are the prime considerations there. Indeed, the work systems in these industries are often referred to as "high-reliability" systems.

One of the more curious findings, however, is that there is little evidence of work practices associated with high-performance production systems even in organizations that have production-like aspects. The processing of transactions in the back offices of financial services and related industries, for example, looks very much like an assembly line (more people are employed in these industries than in manufacturing). Yet there appears to be little—if any—evidence that high-performance production practices or even specific high-performance work practices are being used in these operations. Indeed, the effort in these facilities seems to be quite strongly in the opposite direction; to automate employees out of the process altogether.[4]

It is not obvious that there is a common trend in service jobs. In health care, for example, anecdotal evidence suggests that the biggest development has been efforts to deskill jobs along the lines of Taylorism: Many of the simple tasks traditionally performed by nurses are now being transferred to lower skilled workers. In customer contact jobs in retailing and hospitality, there are some efforts to "empower" workers by giving them more authority to solve problems. Overall, there appears to be a clear trend toward high-performance work in production-oriented jobs because it is associated with a new production process. It is not clear that this movement will make the same inroads elsewhere.

What Skills Have Changed?

In situations where new work practices are in place, how have the jobs changed? Consider, for example, the tasks transferred to work teams in high-performance work systems in a manufacturing environment. The systems of performance measurement and control are already in place, as is the existing job design. The task facing the teams is simply to learn how to interpret information from the system in order to look for ways to improve it. They are not designing and setting up a new system. Further, because these decisions are made in teams, it is not necessary for each worker to have all of the skills needed to handle every task, only that those skills be available somewhere in the work group, perhaps spread across different individuals. For example, not every worker in the group needs to understand how to use statistical process control techniques. If one person understands the notion of confidence limits, another can read the charts, and a third knows his or her machine tools well enough to troubleshoot when the problems have been identified, they have a team that can make the technique work.

Another study examines the relationship between these new work practices and skill needs using data on jobs from the public utilities industry (Cappelli and Rogovsky 1993). The workers were asked about the skills they needed to improve performance in their jobs and also about the extent to which they used work practices associated with high-performance systems. The overall results suggest that there are some, although not many, significant differences in skill needs associated with high-performance work. And some of the differences suggested that skill needs were actually lower where there was more high-performance work. For example, skill needs were lower where certain team processes were in place, perhaps because individual workers must function on their own and make more decisions by themselves. As a result, each worker would need more knowledge and skill to perform a given task than when that task is performed in a team where knowledge and skill can be pooled across team members. Overall, the skills that tended to be associated with these new work practices are behavioral skills such as working in teams.

These results suggest that while new work practices may make new demands on worker skills, the demands may not be overwhelming, and they may focus more on behavioral skills than on traditional vocational skills. Thinking specifically about "lean-production" systems in manufacturing, the fact that Japanese auto companies can take inexperienced workers in the United States and in the United Kingdom and produce autos more efficiently than can German companies in Germany where craft work skills are thought to be much higher suggests that the skills required by lean production in particular can be taught relatively easily. New production systems may require learning about concepts such as continuous improvement and statistical process control, but much of the training in Japanese auto companies, in particular, is with these behavioral skills and socialization.

Two other developments related to these trends in work organization are changes in the organizational structure of establishments. The organizational chart that represents the hierarchy inside organizations is getting flatter as the "middle" positions are cut back. The empowerment and team work trends noted above help reduce the need for supervisors, an effect that spills over to higher management (i.e., fewer managers are needed to direct supervisors). New information and control systems automate the compliance functions typically directed by middle managers. And the move toward decentralization—e.g., profit-centered operations—reduces the importance of compliance. Flatter organizational charts mean shorter job and promotion ladders inside the organization. The positions that remain, in turn, become broader.

An overall summary of how work may be changing includes the following conclusions:

- Work practices are changing, with more establishments using teams, employee participation, and other such arrangements. But these arrangements are by no means in all industries and occupations and are not yet close to being a majority. While the prospects for increased diffusion look good, there are also important reasons for believing that there will be limits to the spread of these practices.
- Where new work practices have been introduced, skills appear to be higher, although how much higher is hard to gauge, and the skill demands that have increased seem to focus on behavioral skills.
- With respect to the nature of these new skills, new production techniques like lean production change jobs by broadening them, eliminating certain narrow jobs, and loading their tasks onto others. Teams, employee participation, and the other more popular new work practices often lead workers to move across a much wider variety of tasks that often include supervisory tasks. Behavioral skills and work-based skills in general appear to have become much more important.
- Many of the above changes make it increasingly difficult to use simple occupational titles as a way of identifying the tasks that workers perform. The tasks that a given worker performs are now much broader and more likely to overlap with what workers do. To the extent that workers do have a core set of unique tasks, those tasks may now take up a much smaller proportion of their working time.

Together, the arguments above suggest that there are important changes in skill needs, although they may be less than revolutionary. More attention to measuring workplace skill needs seems to be in order, particularly as they stack up against the skill set that workers bring to their jobs. The fact that job titles may no longer be good proxies for what one does in a particular job argues for direct measures of tasks performed in each workplace setting. Finally, data collection efforts need to pay more attention to behavioral skills as they seem to be increasingly important in the workplace.

CHANGES IN THE EMPLOYMENT RELATIONSHIP

The second, related work force development is a breakdown in the traditional relationship between employer and employee. The declining obligations and commitments that employers have, especially for their white-collar workers—and the reciprocal decline in the commitments of employees—raise some profound questions about how work-based skills in particular will be developed in the future. This development is closely related to the issue of lifelong learning, that is how the need for skills will be met once workers are in the labor force.[5]

The circumstances that helped create formal arrangements for managing employees in large firms, often referred to as internal labor markets, are changing. Internalized employment arrangements that buffered jobs from market pressures are giving way to arrangements that rely much more heavily on outside market forces to manage employees. There are a number of reasons for that transformation. They include increased competitive pressures on costs and from investors, especially institutional investors, who are demanding higher profits from publicly held enterprises. In addition to the pressures on costs, another factor associated with changing product markets is the need to react quickly to changing consumer demand. The flexibility required to adapt to changing product markets means that fixed costs, including the fixed costs of internalized training and employment systems, become more difficult to support financially. Public policy also contributes to the breakdown of traditional employment relationships. As the legislative protections on regular employees rise, the administrative costs of using such employees rise as well, especially as compared to using contract workers or temporary employees.

Perhaps the most compelling evidence of the changing employment relationship is the decline in job security. One aspect of this change is the continuing pace of downsizing, which appears to actually have increased through the 1990s even as the economy improves. Econometric evidence suggests that the displacement rate for prime age men (35–55) has doubled in the 1990s as compared to the 1970s (Medoff 1993). Employee tenure with their employers' also appears to have declined, especially for older, white men, the demographic group traditionally most protected by internal labor markets. Most important for the discussion here, attachment to one's occupation is actually increasing even while tenure with one's employer is declining (Rose 1995).

The fact that people are staying in the same occupation longer means that there is a greater incentive for them to invest in occupational training because there is a longer time period in which it can pay off. Yet the fact that tenure is declining implies that there is less incentive for employers to provide that training because the contribution from the employee will be made over a shorter period.

The evidence on changes in training is mixed. There is considerable evidence that new work systems demand new and different skills from employees and that employers who are introducing those systems must train employees to function in them (Osterman 1995). And there is some evidence that this type of training—to improve one's job skills in one's current job—is provided to more workers now than in the past (although the intensity of training appears no greater). But training to learn new jobs has declined compared to earlier periods (Constantine and Neumark 1994).

Many other changes suggest how the attachment between employers and employees may be weakening. The use of temporary employees, for example, has increased by a factor of three since 1985. Even wages exhibit the changing relationship. The returns in the form of higher wages associated with longer service with the same employer have declined sharply over the past decade. Conversely, the costs of changing jobs has virtually disappeared. In the 1980s, for example, workers who changed jobs every other year saw almost the same earnings rise in the late 1980s as did those who kept the same job for 10 years (Marcotte 1994). Several studies report that the pay practices inside firms are now much more subject to market forces than in the past. One particularly striking aspect of that change has occurred with respect to pensions and retirement benefits. In 1979, 83 percent of all the workers who had pensions had defined benefit plans where the benefits were guaranteed and the employer took the risks associated with funding them. By 1988, the most recent data available, finds that figure falling to 66 percent. The change has been due to the growth of defined contribution plans like 401(k)s where benefits are no longer guaranteed and the employees take the risk of maintaining their benefits (Ippolito 1995). Further, with no vesting requirements and no fixed pension costs, these new arrangements create no incentives on either employees or employers to stay together.

The breakdown of attachment between employer and employee raises a number of issues that, in turn, have implications for data collection. Perhaps the most important is the question of how skills and training will be acquired. If workers move between employers more frequently, then the ability of employers to fund training for these workers decreases, at least relative to the demand. Workers are increasingly expected to manage their own careers and seek out training themselves to improve their skills. Especially if workers are staying in the same occupations longer, they are more able to reap the gains of improved skills. We should expect much more of a market to develop for training as workers look outside their current employers for training.

As workers move from employer to employer, we might expect them to stop at schools in between to upgrade their skills. Here the notion of lifelong learning has some powerful policy relevance as the demands on schools will change. In terms of data needs, it is important to learn what these returning workers will demand from schools by way of upgrading their skills; for example, what kind of work experiences create what skill needs at which point in one's career? What makes some workers come back to postsecondary institutions while others go to vendors or alternative providers?

Markets require information. In this case, the labor market will require more information on the skills that workers have as they change jobs, and employees will want to know both what skills are required in different settings and where they can go to get those skills. We might expect greater data needs both from and for all three groups—employees, employers, and schools.

One way to think about this new situation is that it may repeat the school-to-work transition problem several times over a worker's career. All the issues about how to make learning more responsive to workplace needs, how to signal skills to employers when leaving school, and so on, get compounded when one is going back-and-forth from school and work.

FUTURE DATA ISSUES

The developments outlined above serve as background to some long-standing questions for which additional education and work-related data are needed. These questions are organized into two major headings:

What Does Work Demand from Employees?

What knowledge, skills, and abilities (KSAs) are required by people entering the work force or already in it that could be met by the educational system, broadly defined? This seems like a unnecessarily general question, but it helps to set up the choices that must be made by policymakers in defining data collection and research questions that can be tracked more easily.

Perhaps the first choice is what does it mean to say that work "requires" something from employees? Does that mean, for example, the requirements needed to get a job—the type of KSAs typically found in job descriptions like those in the *Dictionary of Occupational Titles*? Such requirements can be thought of as either the minimum needed to carry out a job or to be competent at it. Or does it mean the KSAs "required" to excel in a job, associated with improved job performance? The two may be very different and not necessarily be matters of degree. Excelling at a job, for example, often means finding ways to go beyond the current standards as defined by job descriptions or finding ways to alter the task requirements.

The minimum competency approach is not really an empirical research question in the usual sense. It is not, for example, derived from the actual experience of employees. Rather, it is more a deductive process based on the a priori requirements as articulated by industrial engineers who design the jobs. Job analyses in personnel psychology essentially collect this kind of information. The analysts ask either experts or sometimes the employees themselves to identify the tasks that they perform and then use various taxonomies to organize the requirements into KSAs. Some of the taxonomies are organized around the traits that employees need to do the jobs, while others are organized around the characteristics of the tasks themselves. The skills generated by the SCANS Commission are based on job analyses that mix the trait and task approach.

Most of the research on whether skill requirements are changing have been based on job analysis-type data like that contained in the *Dictionary of Occupational Titles*. It is important to understand what exactly such measures can tell us. They capture a point-in-time assessment of what employers ask employees to do with respect to the organization of work. They do not attempt to assess whether what they are doing makes sense and whether it in fact contributes to performance. For example, a job analysis of manufacturing jobs 10 years ago would reveal a set of required KSAs (e.g., emphasizing compliance and downplaying initiative) that now are seen as retarding improved performance in the light of "high-performance" work organization in manufacturing that is both dramatically different and apparently much more efficient than in the past.

Job analysis data might therefore not be especially valid as an indicator of what skills are really needed in the future. What employers are doing at any point may not be optimal and in any case is always likely to change. (Many observers suggest that we have a skills problem in the United States precisely because we set out expectations for the educational system based on what employers demanded from front-line workers 10 years ago, which was very little.) Job analysis data over time might be a better indication about the trends on how employer requirements have changed.

A related use of job analysis-style information is to estimate how changes in the distribution of employment across occupations may affect future skill demands in the economy as a whole. For example, a shift in employment from manufacturing toward clerical jobs means that the skills required in the average job will change. But the problem noted above still applies: Current skill requirements of jobs may not reflect optimal or even future requirements.

Validating job analysis data is problematic without some other independent set of information on job requirements. More to the point, requirements from job analysis data are rarely related to actual job performance measures. Again, job analysis data indicate only what is required for minimum performance and do not suggest what KSAs are required for superior performance. It could well be that the KSAs required for superior performance in a job are very different from those described by job analyses for minimum competence. The way to tell, of course, would be to examine the relationship between KSAs and actual job performance. Such relationships answer a different question—what predicts better performance? The ontology behind this approach is very different than that described above. While job analysis is a kind of deductive process where a given task is mapped onto KSAs using a set of established algorithms to identify job requirements, real validation efforts reveal underlying relationships between KSAs and performance by looking for statistical relationships. There is no reason to expect that the two approaches will yield the same results.

The validation approach of comparing actual job performance to worker characteristics has several important advantages as a means for identifying the KSAs that are important for work. First, it does not require algorithms or judgments about linking tasks to KSAs. Nor does it require mapping out what an individual employee actually does on the job. As noted earlier, identifying the full range of skills one performs on the job becomes increasingly difficult as jobs become broader, and more flexibly defined, and workers are given substantial autonomy over both what tasks they perform and how those tasks are carried out. As noted earlier, what an individual actually does in a particular job title may well vary day-by-day now as well as by situation (e.g., two secretaries with the same job title may do very different things depending on who their boss is).

Further, the validation approach of looking at actual performance makes it much easier to see relationships with educational characteristics. With job analyses, the particular set of KSAs being labeled varies with the type of job analysis chosen. And mapping a given taxonomy of KSAs onto educational characteristics is not at all straightforward. For example, if a job analysis reports that a given job requires a high level of problem-solving skills, what does that say about educational requirements? Does it mean that graduates will do better with more math or logic courses, or is the problem-solving so contextually oriented that something like engineering courses are really what is required? The validation approach would provide direct answers to these questions by showing the effect of different course-taking patterns on student performance.

Job Analysis Data

The National Job Analysis Study currently being undertaken by American College Testing represents what will be the best information available on current job requirements for the economy as a whole. It is designed to provide something like minimum competencies for broad clusters of jobs across the economy as a whole. In terms of additional data collection in this area, the most useful approach would be to repeat something like this study at a later date in order to assess whether these average competencies are changing—not only whether employment shifts across occupations are affecting average skill levels but also whether the skills of particular occupations are changing.

Beyond the job analysis-style assessment of average competencies, which are essentially impossible to validate, it is less obvious how this job analysis data can be used. It will represent something like a taxonomy of relevant skills that has been grounded in field-based experience. Not all of the skills it identifies will be relevant for education, however, as some may be quite job- or context-specific. Most observers would agree that the focus for education should be on the KSAs that are at least to some extent cross-functional, extending beyond individual jobs and, at a minimum, onto careers within general occupational areas. Determining how many KSAs are truly relevant across all jobs is a difficult question, and whether policy makers want to focus down to the level of specific occupations, losing generality in the process (as the National Skills Standards Board is doing), or aggregate up to some higher level, thus losing specificity, is a difficult choice.

The skill information from the National Job Analysis Study can also be used as a taxonomy for collecting further information on job requirements. For example, if it turns out that certain skills feature prominently across occupations in the job analysis data, then perhaps we need to collect data on those skills—e.g., how widespread they are—for other analyses.

The first issue might then be which skills to include. The distinctions used in the *Dictionary of Occupational Titles* between basic, cross-functional, and occupation-specific skills seem to be the most appealing criteria to use as a way of including skills into a classification scheme. They strike a reasonable trade-off between parsimony and richness and get at the kind of information that is relevant in the labor market. Campbell (1994) offers a good assessment of what is required to make such an arrangement work.

But collecting data on the KSAs relevant for education is a problem. Stevens (1994) and others have raised the important practical issue of the limits imposed on any classification system when it goes into the field. The issue of parsimony needs to be considered from the perspective of the NCES operations that are compiling the data. For the reasons noted above, it is unlikely that simply asking a respondent's occupation will provide accurate information about what he or she does on the job and what skills are needed. Many more detailed questions are required, but a population survey has a fixed and relatively small number of questions it can ask.

Consider the current arrangements at the Census, for example. The Current Population Survey (CPS) asks respondents about their business or industry, the kind of work they do, and their most important activities at work (Census 1989). This is not a great deal of descriptive information about the job. Classification clerks then take these responses and aggregate them into occupational codes. In about half the cases, employees believe that their occupation is something different than does their employer (Mellow and Snider 1995). At least half the time, then, one of the parties—employer or employee—is wrong in labeling an occupation.

In other data collection efforts, respondents give the interviewer their job title. Dempsey (1993) suggests that about 10 percent of employers participating in the Department of Labor's Occupational Employment Survey simply submit their current job titles for Census data collection efforts. Researchers then use information from the D.O.T. or other sources to infer information about what skills are required for that job title, ultimately generating estimates for the sample about skill requirements and other issues. The problem, of course, is that the job title the respondent has in his or her organization may be idiosyncratic. It may not correspond well at all to the title that someone in another organization doing the same tasks may have. As noted above, organizations may be getting more idiosyncratic in their job titles, making it even less desirable to let respondents classify themselves.

Interviewers really need to ask respondents directly about their jobs in order to get detailed information on tasks and skills. The experience in Ohio suggests some lessons for how a data collection system might be implemented. Somers (1993) reports that the Ohio Bureau of Employment Services resorted to a series of keywords and computerized text searches for matching workers with jobs, adopting aspects of the Canadian JOBSCAN system for mapping work-related skills that rely on simplified checklists, like keywords, which can be updated easily as jobs change. Perhaps it is possible to use simplified taxonomies like these for measuring the skills required in jobs.

It is important to remember, however, that all of this information is still only about *jobs*. It reflects only minimum requirements of the kind described earlier and cannot be used for any validation efforts relating skills and performance. That requires collecting data on the KSAs individuals possess and then comparing them to some measure of actual job performance.

What Predicts Workplace Success?

As noted above, job analysis-style information that establishes minimum competencies is not the same thing as identifying success on the job. Efforts to identify the characteristics of workers that predict labor market success, almost uniformly defined as wages by labor economists (sometimes unemployment or other labor force status measures are used as well), explain relatively little of the total variance in the outcome or success measure; in fact, they explain rarely more than about a third. Personnel psychologists generally use broader, but potentially more subjective, measures of job performance such as the evaluations of supervisors. Their efforts at predicting performance are more successful, sometimes explaining as much as half of the variance in outcomes, but the studies have other methodological drawbacks such as non-random selection.

One of the most basic needs for research is simply to provide some validation on the basic issue of what work demands from employees in terms of KSAs by relating those KSAs to actual job performance. Once we have job analysis-style data, can we show that those KSAs in fact predict an individual's job performance? That need, in turn, makes some important demands on data. The first, as noted above, is simply to measure the relevant KSAs in employees. This demand leads to an important question: What is the boundary between KSAs obtained from education and from other areas?

The KSAs that are presumably of greatest interest to NCES are those that are related to educational institutions, those that one would expect to be learned in schools. But in practice, the KSAs relevant to success in the workplace are likely to be learned in the family, in school, and in a wide variety of settings that are difficult to separate. This is especially the case where school-to-work programs have been introduced with the goal of blurring the distinctions between these categories of learning.

One approach to this problem is simply ignore it, and to rely instead on traditional measures of academic achievement that measure classroom learning. School-based credentials like degrees, grades completed, and achievement test scores measure what has been presented to students in the school setting. No doubt they are unlikely to represent all or perhaps even most of what is relevant to workplace success. But when related to measures of such success, they do allow one to address whether education matters for workplace success and, if so, which aspects matter. This is obviously more limited than knowing what workplace success demands in terms of KSAs. But knowing how traditional academic achievement matters for workplace success would still be a considerable achievement over where we are now.

Within the general heading of understanding how educational experiences affect employment outcomes are three subquestions:

Better Data on KSAs

Perhaps the first question is simply to develop a better understanding as to what education-related characteristics, or KSAs, determine how well a student does in the labor market. The place to start is to get better information on what the components of an individual's KSAs might be. As noted earlier, traditional measures of academic achievement help us understand how student achievement in the context of current curricula and pedagogy affect labor market success. But this is still a bit of a black box in that we cannot unbundle the subcomponents of academic achievement. For example, if grade point averages predict job success, is the power of the grades coming from the academic knowledge they measure, the comportment aspects they capture (attendance, perseverance, and so on), or the more general problem-solving and organization skills that help determine academic success?

Within the context of academic success, we first need better measures of academic achievement that go beyond traditional grade point averages. The data sets that include standardized test scores are clearly an improvement over grades alone in that they allow us to measure cognitive performance independent from the classroom experiences that affect grades (attitudes, participation, and so on). Several NCES data sets already include such measures. Including more general cognitive ability tests like the General Abilities Test Battery (GATBy) in data also captures something different from subject-based achievement tests. These measures have contributed in important ways to research on labor market outcomes (Tyler, Murnane, and Levy 1995). One problem with such tests, however, is that they tend to be unreliable unless students have a real stake in doing well on them; tests that are administered simply for the purposes of the survey will find students not making the effort to do well on them, thus biasing the results. It is not obvious how to address that problem, which means that samples using such tests will have important biases (either they exclude those who do not take them, a group that is systematically different in other ways, or they include them and somehow try to account for the fact that their performance will be worse).

Currently, one of the most fundamental questions in the topic of employment is the extent to which job performance is driven mainly by cognitive ability, as some have argued (Ree 1994). If this is so, then perhaps curriculum and pedagogy should be redesigned to emphasize cognitive development. But we need better data and more research to identify whether this really is the case. For example, the data used to argue for the importance of cognitive ability in personnel psychology typically do not include measures of an individual's educational experiences; therefore, it is impossible to tell whether the measures of cognitive ability in fact stand as proxies for aspects of education that covary with cognitive ability.

It is also clear, however, that a wide range of important educational experiences are not examined by current data. Extracurricular activities, for example, appear in the research noted above to be very important in shaping workplace performance but are not typically measured in any detail in current surveys. Particularly with regard to the transition from school-to-work, some of the most important experiences facilitating that transition may take place outside of school. And while basic information on work experiences is currently collected in several NCES databases, it would be helpful to have more detailed information on what actually happens to student workers in the workplace. For example, how are they supervised? Do they receive any formal or informal training and, if so, of what kind? What is the nature of the tasks that they perform? Questions like these are very important in understanding what helps students make the transition to the workplace and in designing curricula to facilitate that transition (see below).

More generally, work-based skills and competencies are not directly measured by any of the national probability datasets, nor are behavioral skills or dispositional characteristics like personality that both prior research and commentary suggest are crucial to job success.

The term "behavioral skills" is a code word for a range of knowledge about issues such as group and individual behavior, interpersonal and self-management skills, and attributes and abilities. The first problem with collecting data on behavioral skills, indeed on any work-based skills, is how to measure them. There are a number of competing taxonomies for such skills like the trait-based job analyses in personnel or the SCANS skills used in public policy. Every taxonomy "cuts" the KSAs in a slightly different way.

The problem for NCES in collecting data on work-based and behavioral skills is first to choose a taxonomy for measuring those skills. The key issue is to choose a taxonomy that does not leave anything out and that avoids lumping important concepts together. The SCANS skills, for example, seem to put together many distinct behavioral skills into the same categories (e.g., self-management and interpersonal skills), making it difficult to interpret relationships with those measures. It might also be important to anticipate which of the various taxonomies will come to be accepted in future policy discussions. Will American College Testing's National Job Analysis Study, for example, be embraced by the research and policy communities, and should NCES use its taxonomy of skills for collecting data on work-based skills? One sure bet is that no single taxonomy will be embraced by the research community. There have been decades of debate and contention regarding the appropriate methods for doing job analyses with no clear consensus emerging as to the "best" taxonomy, because each represents trade-offs on issues about which reasonable people can and do differ.

Perhaps the best advice on this issue is to have the various government agencies interested in measuring work-based skills agree on a taxonomy and get on with it. Objections will be raised no matter what is chosen, but if there is agreement among the government players, the taxonomy selected will become the standard: "If you collect it, they will use it."

How to measure work-based skills, particularly behavioral skills, is a more complicated problem. It may be possible to proxy skills with certain credentials like coursework related to behavioral skills. While taking a course in interpersonal skills may not seem like a good proxy—indeed, it may simply select in those people who have bad skills and are taking it because they really need help—the same procedure is generally used to measure one's academic skill base in a subject area like math. In the absence of clear credentials, it becomes difficult to rely on self-reporting, and surveys must find some other way to measure skills. In the area of academic achievement, a series of well-established standardized tests are available for measuring subject knowledge and various abilities. There are no real equivalents yet on the behavioral side, although there are well-accepted tests in specialized areas like personality profiles. But someone will certainly seize the enormous opportunity that tests of behavioral skills offer in improving employee selection, and those will soon be available.

Better Measures of Education Institutions

If we had a better understanding of which student characteristics lead to success in the workplace, it would then be important to learn what characteristics of educational experiences, broadly defined, help produce those characteristics.

The "toe-in-the-water" approach to additional data in this area is to collect further, more detailed data on classroom experiences. Most of the research on education and labor market outcomes has been limited to looking at gross measures of educational attainment—years of education completed and degrees conferred. Perhaps the most important innovation in contemporary research has been to add detail to those existing measures. The NCES data on student transcripts, for example, has made possible new research on the effects of patterns of course taking on labor market outcomes (Altonji 1995). This research has been well received and has already contributed in a central way to policy debates such as the relative returns to attending

2-year versus 4-year institutions (Kane and Rouse 1995). What is perhaps most surprising about this line of research is how long it has taken to get started and how much remains to be done. It is possible to count almost on one hand the number of studies that have looked at the content of student coursework as it affects labor market outcomes.

A few studies in personnel psychology have explored the impact on job performance of student experiences in addition to course-taking patterns. These include, for example, studies of extracurricular activities where the results suggest that these experiences are very powerful predictors of job performance, more powerful, in fact, than academic performance (Bray, Campbell, and Grant 1974).

A related development, also in its infancy, has been to look at the characteristics of educational institutions as organizations that affect the labor market performance of their graduates/attendees. There are many studies that look at how the characteristics of postsecondary schools and teachers affect the academic achievement of their students (see Hanasheck et al. 1994 for a recent review), but again, very few that link those characteristics to labor market outcomes. For example, no studies have looked at the relationships between aspects of how schools are organized and the labor market performance of their students (Johnson and Summers [1993] review this literature at length.)

Among the very few studies that attempt to link school characteristics to labor market outcomes of their students are Crawford, Summers, and Johnson (1994) for secondary schools, and Daniel et al. (1995) for higher education institutions. The results suggest that the characteristics of these institutions do matter, but the measures are aggregated at a level that makes it difficult to see relationships with specific practices and to offer detailed guidance on organizing schools.

The data problems in linking school characteristics and labor market outcomes begin with the fact that most of the surveys that collect longitudinal labor market data are national probability samples where it is unlikely that many respondents will come from the same institutions. The pathbreaking analyses will be to look within institutions to see how variations in education experiences affect student performance—both traditional academic achievement and labor market outcomes. To illustrate, data that might find better student performance associated with attending small liberal arts colleges is confounded: Is the better performance the result of smaller class size, small academic communities, the typical liberal arts curriculum, or the characteristics of students selected into such schools? We would need to look at the variance in experiences within these schools in order to answer those questions.

The data required to address these within-institution questions are considerable: first, the data must be longitudinal, following students through their postsecondary experiences and into the labor market; second, they must represent samples of reasonable size within postsecondary institutions; and third, they must include a wide range of such institutions. These data needs are considered in more detail below. If available, they would offer an enormous research opportunity for relating traditional measures of academic achievement and school characteristics to labor market outcomes.

Better Data on Work Outcomes

The arguments above suggest the need for better information about the knowledge, skills, and abilities that individuals possess in order to explain work outcomes and, in turn, determine what KSAs are really demanded in the workplace. Even with this better information, however, there is a weak link in the analysis, and that is the measure of workplace outcomes and performance.

As noted earlier, the majority of studies relating education and work outcomes use wages as the measure of "success" on the grounds that superior performers will be rewarded with higher wages, other things equal. But there are some obvious difficulties with that approach. For example, wages are driven perhaps most strongly by occupational choices and not performance within an occupation; the best school teacher in the world still earns less than a mediocre investment banker. Occupations differ greatly in how wages relate to performance. A good sales associate may earn substantially more than a poor one, but a good teacher is likely to earn about as much as an average teacher. In general, the relationship between performance and compensation may not be especially strong across the economy.

It is certainly possible with modern econometric techniques to address some of these issues. For example, looking at wages within occupations, controlling for employment status (i.e., wages conditional on having a job and on working hours) and other factors that might affect pay, may address some of these issues. But short of perfect modeling, these are at best imperfect adjustments. For example, someone who pursues his or her occupation in the non-profit sector of the economy will earn less. The characteristics that lead someone to make that decision (e.g., attending a college with public service requirements) will turn up in a validation exercise as being negatively associated with earnings and, in turn, appear as something that actually hinders workplace performance.

Some improvement comes with expanding the range of labor market data on individuals to include, for example, spells of employment, long-term career earnings, training received and career mobility, job and life satisfaction, and so on. Ultimately, however, we need better information about the nature of work performance for individual workers.

Specifically, it would be important to know not only whether a worker is doing well or not but also which aspects of their performance are good and which are poor. Ideally, we would like that information in ways that tie directly into KSAs—are there skills that the employee seems to lack, for example, that are associated with poor performance? Such information would be especially helpful to know for new entrants/school leavers where the link between education and performance may be most clear. There is a perception, for example, that the school-to-work transition problem is in part due to comportment problems and poor self-management skills among school leavers. Detailed information on their performance would be especially useful to address that issue.

A survey conducted by the National Foundation of Independent Businesses (NFIB) offers one example of alternative performance data on employees. The survey of employers asked a series of detailed questions about the last employee hired and his or her job performance (actual versus expected). The Department of Labor in the State of New Hampshire collected similar data on school leavers by going to their employers and asking detailed questions about how those individuals were performing in the workplace. Personnel psychologists routinely collect such data on a wide range of performance outcomes, including promotion potential, organizational citizenship, and so on.

The main difficulty with alternative performance data is in collecting it. Unlike wage data, these data cannot be self-reported accurately, and many questions must be used to produce reliable scales for each concept. Such data must be collected from employers. Surveys like the General Social Survey and the National Organizations Survey have collected matching data from employers and their employees by asking respondents to identify their employer and then contacting and surveying the employer. The additional problem with individual performance measures is that it is unlikely that a centralized personnel office could complete surveys about aspects of a specific employee's performance, especially in large establishments. Supervisors within the establishment may have to be enlisted to answer the questions, raising rater reliability issues and reducing the expected response rates. When personnel psychologists collect such data, it is typically within a single organization where the organization's own performance measures can be used. These may be consistent within that organization, but they are unlikely to be consistent across different organizations.

Work Performance Beyond the Individual

As noted earlier, the interest in how education affects workplace performance has been driven not just by the belief that it might improve an individual's performance and earnings but also by the view that it might make both establishments and economies more productive and effective. Research such as that performed with the National Employer Survey (EQW 1995), which finds that establishments with a more educated workforce, other things being equal, are more productive, has been the focus of considerable policy interest.

How NCES might develop data to expand the measurement of performance is worth considering. The first issue to confront is that it would require performance-based information on groups larger than individuals—teams or work groups, establishments, and so on—an effort that might seem far beyond the traditional paradigms of NCES data collection. But there are some exceptions even with the data that NCES already collects. For example, it collects detailed organizational information on one type of operation; schools, using the Schools and Staffing Survey. Studies examining how the educational background of school staff affects student performance are already relatively common. It would not require much new data on the educational experiences of teachers and administrators to examine the relationships between establishment-level performance and the particular experiences of school staff.

Beyond this education-specific setting, there may be ways to join forces with other establishment-level surveys in order to examine the performance effects of education.

How Does Work Affect Education?

While most of the recent policy interests seem to be focused on the question raised earlier of how education can contribute to workplace success, the more traditional and equally important question is how work experience affects academic achievement. How secondary school work experience affects students' educational performance is a question with a significant research tradition, but several more contemporary issues also demand attention.

Given that so many students work while attending school and the trend toward combining work and school in postsecondary education seems to be increasing, it is very important to know how traditional work experiences (i.e., part-time jobs) affect educational performance. We need to go well beyond existing research, which has focused mainly on how hours of work affect student classroom achievement, to understand how the characteristics of that work experience affect academic performance. The general public understands that the nature of the work experience is crucial to educational success, as evidenced by the different language we use to describe different student work experiences (i.e., internships versus part-time or summer jobs). Consider some of the following research questions:

- Especially for secondary school students, what effect does working in a stereotypical fast-food or low-skill job have on academic performance? When, for example, student workers are often supervised by school dropouts barely older than the students themselves, are there negative "modeling" effects that lead to worse academic achievement?

- Especially for postsecondary education, does having a "better" job that offers more opportunities for learning and advancement while attending school actually contribute to dropping out as employers pull the best students out of school and into full-time jobs? Or does it allow more students to complete school by increasing their resources? Does it change their course-taking patterns and choice of major? Do students with more work experience have a smoother transition to the workplace after graduation?

- What effect does work experience have on KSAs other than the classroom-based knowledge measured by traditional achievement tests? Do different kinds of work experience provide alternative vehicles for learning SCANS-type skills, for example?

- How do different kinds of work experience affect postsecondary school experiences—attendance, completion rates, course and major selection, and so on?

- Finally, how does work experience shape the demand for continuing education? Do different kinds of work experience make it more likely to pursue postsecondary education? For example, does a part-time job in a hospital, where one learns about all kinds of careers that require further training, make one more likely to pursue further education than if one did the same kind of unskilled work (e.g., janitor) in a different setting? Even for students who do not attend traditional postsecondary institutions, do different kinds of work experience make them more likely to pursue skills and training through other avenues?

In Secondary School

Researchers have argued back and forth about the effect on student achievement and subsequent educational plans of working while in school. With few exceptions, this research has focused on the quantity of work, with relatively little attention paid to the quality of the work experience. As argued above, better information on the characteristics of a student's working experience would help considerably in understanding the real impact on education. Such information and data are a special priority at present given the introduction of school-to-work transition programs across the country and the need to understand what makes them successful.

The type of evaluation of vocational education programs recently conducted as part of the legislative reauthorization would also be enhanced considerably by knowing the characteristics of the work experience in those programs. It might well be, for example, that there are no real differences between youth apprenticeship programs and cooperative education programs and that the apparent variance in their results is simply due to the characteristics of the work experience in each setting.

In Postsecondary School

All of the above issues apply to student experiences in postsecondary school as well, although they have been far less researched. Student working hours and experiences may have important impacts on academic performance as well as various kinds of institutional arrangements such as co-ops programs and summer internships. Whether and how much students work in school is linked closely to issues of student financial aid and school resources, another important policy issue.

Lifelong Learning

The issue of education after entering the labor force needs to be put squarely on the research agenda. As the length of time many students attend postsecondary school gets longer and increasingly is combined with full-time employment, it no longer makes sense to think of this as simply delays in graduation. It may be more appropriate to think of this situation not as a transition period to graduation but as a new and stable pattern: going back and forth from work to school, taking new courses as workplace demands require them, and possibly making career and work changes as new skills are acquired. All of the above issues as to how work experiences shape educational choices and outcomes apply to these new "lifelong learners" as well.

RECOMMENDATIONS FOR DATA

Most of us would be delighted to see NCES develop new data sets specifically tailored to meet some of the concerns noted above, but given the tremendous investments required for such efforts, it would be impractical at best in the current climate of fiscal restraint to make such recommendations. In fact, some of the important questions can be addressed using existing data, and relatively simple additions to the data series currently maintained by NCES would address many of the remaining data needs.

The most basic data need is to have information in the same data set about an individual's educational and work experiences. An issue that is integral to many of the more specific questions raised above is simply to get a better understanding of what demographers refer to as the "life course" of young people. Have the paths from school-to-work or secondary to postsecondary school changed? Consider some of the basic factual questions embedded within that more general question for which we currently do not have good answers:

- Are more postsecondary students working full time?
- Has the pattern of "articulation" or transfer of students from less-than-4-year to 4-year institutions changed?
- Are postsecondary school graduates returning to school after entering the work force to upgrade their skills?
- How many secondary school students participate in school-to-work programs?

NCES already maintains a number of data series on individuals and their educational experiences, as shown in Figure 1.

Figure 1—Availability of data in NCES sources that can be used to measure components of school-to-work

Type of data & school-to-work components	HS&B	NELS	NPSAS	BPS	B&B	SASS
Type of data source	longitudinal	longitudinal	cross-sectional	longitudinal	longitudinal	cross-sectional
Years of collection	1980–86, 1980–92	1988–94	1987, 1990, 1993	1990–94	1993–95	1988–91, 1994
Level at which data are specified	student & school	student & school	student & institution	student & institution	student & institution	school
School-to-work components						
A. Educational preparation for work						
1) Educational Attainment	yes	yes	yes	yes	yes	NA
2) Postsecondary enrollment & persistence in school	yes yes	yes yes	yes short-term only	yes yes	yes yes	NA NA
3) Transcript data	postsec. only, sec. & postsec.	secondary	no	no	forthcoming (postsec. only)	NA
which can be used to:						
• distinguish among students with similar degrees	yes	yes	no	no	forthcoming	NA
• measure attainment	yes	yes	no	no	forthcoming	NA
• specify a career major	yes	secondary only	no	no	postsec. only	NA
• assess exposure to all aspects of an industry	yes	yes	no	no	postsec. only	NA
4) Grades	yes	yes	student report	student report	forthcoming	NA
test battery scores	yes	yes	no	no	no	NA
which can be used to:						
• develop gain scores	yes	yes	no	no	no	NA

Figure 1—Availability of data in NCES sources that can be used to measure components of school-to-work—Continued

Type of data & school-to-work components	HS&B	NELS	NPSAS	BPS	B&B	SASS
B. Work experience						
General availability of measures						
Employment Status	monthly 1980–86, 1982–92	monthly 1992–94	annualized	monthly 1990–94	monthly 1993–94	NA NA
Wages	1980–86 only	yes	limited	yes	yes	NA
Earnings	yes	yes	yes	yes	yes	NA
Avg. hours per week	1980–86, 1982–86	yes	yes	yes	yes	NA
Occupation	yes	yes	no	yes	yes	NA
Industry	yes	yes	no	yes	yes	NA
Relatedness of employment to education[2]	student report & linked codes	student report & linked codes	student report & linked codes	student report & linked codes	student report & linked codes	NA
Availability of measures by topic						
1) Employment experiences in high school	yes	yes	no	no	no	NA
2) Employment exp. in postsec. enrollment	yes	yes	limited	yes	yes	NA
3) Employment exp. as an outcome	yes	yes	no	yes	yes	NA
C. Patterns & processes of articulation						
1) Secondary to postsecondary	yes, with 10 yrs. post HS	yes, with 2 yrs. post HS	no	no	no	NA
2) Postsecondary to postsecondary	yes	no	no	yes	perhaps w/transcripts	NA

Figure 1—Availability of data in NCES sources that can be used to measure components of school-to-work—Continued

Type of data & school-to-work components	HS&B	NELS	NPSAS	BPS	B&B	SASS
3) HS or postsec. to employ.	yes	yes	no	yes	yes	NA
D. Availability of institutional resources						
1) Number of HS w/work prep. programs	yes, but dated	% of students in programs	no	no	no	yes
2) Number of postsec.inst. w/work prep. programs	no	no	yes	yes	yes (BA/BS only)	no
3) Availability of teachers to teach integrated academic & applied curricula	no	perhaps, but not representative	no	no	no	yes
Background items Student characteristics	yes	yes	yes	yes	yes	no
Family characteristics	yes	yes	yes	yes	yes	no
School or institutional characteristics	yes	yes	yes	yes	yes	yes
Community type	yes	yes	no	no	no	yes
Attitudes and expectations	yes	yes	some	some	some	no
Population characteristics	yes	yes	yes	yes	yes	yes

NOTES: "Yes" indicates that the data set includes items in which the school-to-work element can be measured; "No" indicates that the database does not contain such items; and "NA" means not available. Other entries indicate that the topic is covered by items in the data set, but that coverage is limited as described.

SOURCE: Medrich, E. and Tuma, J. *School-to-Work Data Available in NCES Data Sources*. 1995. Washington, D.C.: National School-to-Work Office.

One can see even from this brief description how rich many of these data sets are in terms of information on education. Several of the data series, like High School and Beyond, the National Educational Longitudinal Study, and the National Longitudinal Study of the Class of 1972 involved collecting data from a respondent's school. Even the richest of these surveys, however, are thin on the following attributes:

- *Content of educational experiences.* Only the three surveys in the above paragraph and the Baccalaureate and Beyond survey have transcript data. And, as noted above, it is difficult to know much about what students actually learned in those courses without more standardized instruments like achievement tests. It would also be helpful to have information on pedagogy—did the classes require written assignments or lab work, was there class discussion or team projects, how big were the classes, were the exams essay or multiple choice? These factors are perhaps even more important to the current debate about education reform than are curriculum issues.

- *Information on relevant KSAs.* None of the NCES data sets currently collects information on behavioral skills or on the kind of work-based skills described by the SCANS report or similar exercises.

- *Details on work experience.* Understanding how work affects education requires knowing about a respondent's work experience. The data currently collected in NCES surveys looks at what might be called outcomes of work—job titles, industry, hours, and wages. What we do not know is what students actually did on the job. What kind of training or supervision did they receive; what tasks did they perform; did they participate in decision making, and so on. As noted earlier, job titles never conveyed much information on these issues, and there are good reasons for believing that they will be even less reliable in the future.

- *Information on job success.* The current NCES data sets have only information about wages and earnings that have limits as proxies for job performance. As noted earlier, it is important to know exactly where workers had success, where they had difficulty, and what skills or tasks were in deficit.

- *Details on employer practices.* If the interest in lifelong learning is real, then it is especially important to know what pushes people back to school after they have joined the work force. The nature of work organization no doubt plays some role in that decision as does a series of employer practices such as tuition reimbursement plans or career planning and progression programs.

Strategies for Collecting New Data

New Data on How Education Affects Work

Clearly the best approach for addressing at least some of the data needs outlined above is to leverage existing data sets by adding data to them. High School and Beyond and the National Longitudinal Survey of 1972 have important attributes in that they contain some reasonably detailed information on education experiences, and, more important, they contain a long enough time series to identify a respondent's long-term job success. Such information is especially important for assessing the effects of education on work. The drawback to such data, however, is that the respondents have typically been away from formal schooling so long that it is very difficult to collect additional information from them about educational experiences.

HS&B and NLS–72 can be supplemented, however, to address some of the questions noted above about the effects of education on work. First, simple questions on job success could add information to the wage data. For example, a few questions asking about job content, a respondent's position in the hierarchy, and mobility would help identify workplace success. Self-reported data on skill needs would be easy to collect. When related to earlier data on educational experiences, these responses would help identify how work affects job success.

These two data sets in particular would be especially useful in addressing some of the lifelong learning issues noted above. Specifically, what makes an individual seek further education, and if he or she does, what kind of education (topic and provider) does he or she seek? Some of the information on educational choices over a lifetime is already in these data sets. What needs to be added are questions about work experiences. First, what is it about the type of work a respondent performs— tasks and job content as noted above—that pushes them to get further education? Is more challenging work the driver, or is it that those who go back for more education eventually get more challenging jobs? Second, what is it about employer practices that encourages lifelong learning? Is it financial support in the form of tuition reimbursement, or is it incentives like merit-based pay and promotion systems? Together with the job performance information above, these new data would allow researchers to know whether lifelong learning contributes to job performance and, if so, the kind of learning and education experiences that affect job and labor market performance.

Several of the problems noted above hinge on getting data about employers such as performance measures that cannot be obtained from surveys at the individual level. Employer-level data is important for addressing questions such as the following:

- How might different aspects of education in a work force affect organizational performance?
- How does having a more educated work force affect how work is organized or other issues of organizational operations?
- What characteristics of employers (and jobs) contribute to increased use of postsecondary education among employees?
- To what extent are specific postsecondary courses and programs substitutes for firm-provided training, especially at the community college level?
- What are the skills that employers demand from their work force, and how might they be changing?

Such information comes from establishment-based surveys like the EQW/Census National Employer's Survey. But NCES does not maintain establishment-level data sets. Such establishment-level data would still leave one with the problem of getting detailed information on the educational experiences of individual employees.

The ideal solution is to provide matching data for employers and employees, asking the relevant questions for each group and then putting the two sets of data together. Two approaches for doing so and constructing sampling frames have been used. The first is to survey a probability sample of individuals, asking them about their educational experiences and so on, and to identify their employers. The next step is to go to their employers and survey them about their practices and performance. This technique was used by researchers conducting the National Organizational Survey (NOS) funded by the National Science Foundation. They used questions from the General Social Survey (GSS) of individuals to identify employers, and the GSS data on individuals was then matched to the NOS data on organizations. For NCES, the best method would be to ask the respondents in existing surveys like NLS–72 and HS&B to identify their employers, survey the employers, and then match the data. One problem with this approach, of course, is that there is only one respondent/employee per employer, and it is very difficult to use the experiences and characteristics of that respondent to generalize about the work force as a whole.

The alternative is to conduct a probability survey of establishments and then survey the employees within that establishment. This is the approach currently being used by the Bureau of Labor Statistics in its training surveys. It is an expensive process, as it requires getting information about the work force from each employer (i.e., the sampling frame) and permission to survey their employees. Another approach under consideration by the EQW/Census National Establishment Survey and used by Statistics Canada in their training survey is to try to survey employees in establishments without knowing the sampling frame in each establishment. But even with this technique, the process is expensive and time consuming. NCES does not have to address every data need itself, and establishment-level data are probably not within its comparative advantage.

The questions noted earlier of relating educational practices at the institutional level to student job and labor market outcomes raise very similar problems for data. Addressing such questions requires matching longitudinal data on individuals and their work outcomes to detailed data on the characteristics of their educational institutions. And many of the same problems of matching individual and organizational data appear here as well; specifically, the need to have many observations from the same educational institutions in order to estimate the effects of within-organizational practices.

Here, the best strategies for data collection do not seem to leverage in any obvious way off of existing surveys. One approach might be to develop a targeted sample of institutions whose education practices and arrangements seemed especially noteworthy or representative, and then to follow a representative sample of their graduates over time to examine their labor market performance. One could then use the data to relate practices and experiences at the classroom level, within institutions, to workplace outcomes.

New Data on How Work Affects Education

Understanding how work affects education is an issue that seems especially within the traditional purview of NCES. It requires information on the nature of work experiences that could then be matched to subsequent education choices and outcomes. The HS&B and NLS–72 data sets discussed above might be used for looking at the effects of work on lifelong learning education choices (e.g., determining who returns to what kind of schooling during their working life). Because the information on working during school is more limited, these surveys are less suited to secondary and more suited to traditional postsecondary education. Such information is best obtained from respondents who are still in school, ideally in secondary school. Existing surveys such as Beginning Postsecondary Students and Baccalaureate and Beyond are missing the secondary school experiences and, as such, are less than ideal.

The best approach is to start collecting data now on secondary school students—or perhaps even students in earlier grades—that will help us to understand how work affects education. Later on, the same data can be used to help understand how detailed educational experiences affect subsequent workplace success. The new data might include the following:

- Detailed information about work experiences during school of the kind noted above including the nature of the tasks performed, type of supervision offered, characteristics of training received, and so on. This information could then be related to subsequent academic achievement, course-taking patterns, and postsecondary experiences.

- More detailed information on KSAs including a student's work-based skills of the kind described by the SCANS report. The idea here would be to see how work experience affects these work-based skills. Later on, such information could be related to success in the workplace to see whether the results are different from those for academic achievement as more traditionally measured.

- Information on school-to-work programs and other work-based learning arrangements associated with schools. What effects do these arrangements have on academic achievement and on subsequent workplace success?

The School-to-Work Opportunities Office has funded two efforts to look at one aspect of these educational practices and arrangements. The first adds questions asking for details of school-to-work programs to the existing superintendent and school administrator's survey administered by the BLS. The second, more relevant here, adds questions to the National Longitudinal Survey of Youth about participation in such programs, information that can then be related to labor market outcomes. There may not be much of an argument for NCES to duplicate that effort with its own surveys. But when any of the existing NCES surveys are again in the field, adding even the same questions on participation in school-to-work programs would enable these surveys to examine the effects that participating in these programs might have on work outcomes. Similarly, the Bureau of Labor Statistics has proposed starting a new longitudinal study of 17-year-olds, and it is possible that this effort may also provide data to address some of the school and work questions.

Finally, there are many ways to collect data for research questions, and surveys of the kind at which NCES excels are obviously only one method for doing so. And it is probably worth a discussion as to what mix of survey data and other research approaches might be appropriate for addressing the questions described below. High-quality survey data with its enormous advantages in external validity are especially useful at capturing main effects of relationships between constructs that can be conceptualized and measured in a straightforward way. It is an important question as to whether selection issues and unmeasured attributes are intractable enough in some topics to demand more sophisticated experimental designs than are provided by national probability surveys. Whether surveys targeted toward particular populations might provide a middle ground between national probability surveys and experiments remains to be seen.

NOTES

1. It is worth pointing out that there is at least as much antipathy on the other side since many employers seem to distrust the goals that educators hold for students ("it's all about self-esteem, the kids aren't learning anything," and so on).

2. See Mishel and Bernstein (1994) for evidence on the former, and Gardner (1993) for evidence on the latter.

3. Much of the material in this section is drawn from Cappelli and Rogovsky (1995).

4. Preliminary findings from a study of transaction processing at the Wharton School's Financial Services Center find virtually no evidence of these practices.

5. These changes are described in Cappelli, P. (Ed.) *Change at Work.* (New York: Oxford University Press, forthcoming). A summary version of the arguments can be found in "Restructuring Employment," *Looking Ahead* (Washington, D.C.: National Planning Association, fall 1994).

REFERENCES

Adler, P. 1993. "Time-and-Motion Regained." *Harvard Business Review* 44(1): 97–108.

Altonji, J. 1995. "The Effects of High School Curriculum on Education and Labor Market Outcomes." *Journal of Human Resources* 30(3): 409–438.

Bartel, A., and Lichtenberg, F. February 1987. "The Comparative Advantage of Educated Workers in Implementing New Technology." *Review of Economics and Statistics* 69: 1–11.

Bowles, S., and Gintis, H. 1976. *Schooling in Capitalist America: Education Reform and the Contradictions of Economic Life.* New York: Basic Books.

Bray, D., Campbell, R., and Grant, D. 1974. *Formative Years in business: A Long-Term AT&T Study of Managerial Lives.* New York, NY: Wiley.

Bureau of the Census. "How to Enumerate the CPS." Washington, D.C.: 1989.

Campbell, J.P. 1994. "The Development of Occupational/Skill Clusters: Issues and Recommendations." Washington, D.C.: U.S. Department of Labor/Education and Training Administration.

Cappelli, P. April 1993. "Are Skill Requirements Rising? Evidence for Production and Clerical Jobs." *Industrial and Labor Relations Review* 46: 515–530.

Cappelli, P. 1995. "Changing Skill Requirements and implications for the Structure of Wages" Paper presented at the Federal Reserve Bank of Boston Conference, November 17, 1995.

Cappelli, P., and Rogovsky, N. 1995. "Skills and Individual Performance." Working Paper. Philadelphia: National Center on the Educational Quality of the Workforce at the University of Pennsylvania.

Constantine, J.M. and Neumark, D. 1994. "Training and the Growth of Wage Inequality." Philadelphia: National Center on the Educational Quality of the Workforce (EQW).

Crawford, D., Johnson, A., and Summers, A. 1994. "School and Labor Market Outcomes." Working Paper. Philadelphia: National Center on the Educational Quality of the Workforce at the University of Pennsylvania.

Daniel, K., Black, and Smith, J. 1995. "College Quality and the Wages of Young Men." Working Paper. Philadelphia: The Wharton School, The University of Pennsylvania.

Demsey, R. September 1993. "An Occupational Classification System for Collecting Employment Data from Both Households and Employers." In *Proceedings of the International Occupational Classification Conference*, pp. 235–250. Washington, D.C.: U.S. Department of Labor, Bureau of Labor Statistics.

EQW 1995. "First Findings: Results of the EQW/Census National Employer Survey." Philadelphia: National Center on the Educational Quality of the Workforce EQW.

Gardner, J. June 1993. "Recession Swells Amount of Displaced Workers." *Monthly Labor Review* 116(6): 14–23.

Hanushek, E. 1994, *Making Schools Work: Improving Performance and Controlling Costs.* Washington, D.C.: The Brookings Institution.

Ippolito, R. 1995. "Toward Explaining the Growth of Defined Contribution Plans." *Industrial Relations* 34: 1–20.

Johnson, A., and Summers, A. 1993. "What Do We Know About How Schools Affect the Labor Market Performance of Their Students?" Philadelphia: National Center on the Educational Quality of the Workforce EQW.

Kane, T., and Rouse, C. June 1995. "Labor Market Returns to Two-and Four-Year College." *American Economic Review* 85: 600–614.

Klein, J. 1989. "The Human Costs of Manufacturing Reform." *Harvard Business Review* 42 (2): 60–66.

Marcotte, D. 1994. "Evidence of a Fall in the Wage Premium for Job Security." Northern Illinois University, Center for Governmental Studies.

Medoff, J. 1993. "Middle-Aged and Out of Work: Growing Unemployment Due to Job Loss Among Middle-Aged Americans." Washington, D.C.: National Study Center.

Mellow, W., and Snider, H. 1995. "Accuracy of Response in Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics* 1(4): 331–344.

Mishel, L., and Bernstein, J. 1994. *The State of Working America, 1994–1995.* Washington, D.C.: Economic Policy Institute.

Murnane, R, Willett, J.B., and Levy, F. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77: 251.

National Commission on Education and the Economy. 1989. *America's Choice: High Skills or Low Wages!* Rochester, NY: National Commission on Education and the Economy.

National Commission for Excellence in Education. 1983. *A Nation At Risk: The Imperative for Education Reform*. Washington, D.C.: National Commission for Excellence in Education.

Osterman, P. 1995. "Skills, Training, and Work Organization in American Establishments." *Industrial Relations*, April, 125–146.

Ree, M., Earles, J., and Teachout, M. August 1994. "Predicting Job Performance: Not Much More Than 'g'" *Journal of Applied Psychology* 79: 518–524.

Rose, S. 1995. "The Decline of Employment Stability in the 1980s." Washington, D.C.: National Commission on Employment Policy.

Sommers, D. 1993. "Ohio's Adoption of Canada's Jobscan Skill Checklists." In *Proceedings of the International Occupational Classification Conference*, pp. 322–335. Washington, D.C.: U.S. Department of Labor, Bureau of Labor Statistics.

Stevens, D. W. April 1994. "Toward the Definition of New Occupational/Skill Clusters for the United States." Washington, D.C.: U.S. Department of Labor.

Tyler, J., Murnane, R., and Levy, F. December 1995. "Are More College Graduates Really Taking 'High School' Jobs?" *Monthly Labor Review* 118(12): 18–28.

Workforce 2000. 1985. Washington, D.C.: The Hudson Institute for the U.S. Department of Labor.

Discussant Comments

DAVID STERN

This is a lucid and lively paper. It builds on the substantial body of original research produced by the EQW Center, of which Cappelli is coordinator and to which he has contributed much impressive research of his own.

If I were to state all the points in the paper with which I agree, I would repeat most of the paper. Instead, I will select a few points to emphasize. And I will express a difference of opinion on one major issue.

Closer Connection Between Learning and Work

Cappelli is certainly correct that education *for* work, and especially education *through* work, have been relatively neglected topics in educational research. If we define education not as schooling, but as intentional learning, then the mere fact that the average person spends approximately 14 or 15 years in school but 40 to 50 years at work, engaging in some degree of intentional learning, should warrant greater attention to education after the end of formal schooling. This is all the more true because the degree to which work involves intentional learning appears to be increasing.

Cappelli describes how "high-performance" or "lean" production have broadened the responsibilities of front-line workers in manufacturing. Production workers have been called upon to learn quality control and job analysis. They are making changeovers to new products and learning new technologies at a faster rate, because their organizations must adapt or die. We are all caught up in accelerating change, born of faster computers, faster communications, faster flows of ideas and capital. This NCES meeting itself can serve as an example of education in the workplace as a response to these changes.

Cappelli also points out that "high-performance" management practices still do not prevail. Rather than investing in education for employees, many employers are choosing to "rent" people instead. The use of temporary staff has tripled since 1985, according to Cappelli. But temporary staff are continually learning, too: they are forced to do so as they move from one job to another. A study of Manpower Inc., the largest of the temporary staffing agencies, indicates that the company helps employees use their experience in a sequence of jobs to build a coherent portfolio of skills for themselves (Seavey and Kazis 1994).

The increasingly educative function of work is evident in the arrangements that some employers have adopted to promote learning. In addition to formal instruction in company classrooms, many firms have devised methods of "just-in-time learning" that minimize the cost of learning by facilitating acquisition of skill and knowledge as part of the work process itself. Examples of these arrangements include cross-training of employees who work near one another, rotation of staff through a planned sequence of positions, and skill-based pay, which compensates individuals in part for what they demonstratably know and can do, independent of their specific job responsibilities during the pay period.

Researchers have debated whether or not changes in the workplace have resulted in a demand for higher levels of skill on the part of workers. Cappelli himself has produced some the most informative studies on this topic. However, as he explains in this paper, the definition of skill requirements is highly problematic. Procedures that personnel departments use to define skill requirements in practice are based on a priori judgments, not on demonstrated empirical relationships between skills and actual performance on the job. Cappelli would like to see more empirical validation studies of this kind. That is one of his main recommendations in his paper. However, I am less optimistic than Cappelli about the feasibility of mapping "KSAs" (knowledge, skills, and abilities) onto job performance. The plethora of distinct KSAs, and the multiplicity of job performance measures, make this research program at least as daunting as mapping the human genome—probably even more so, because job performance depends on contextual variables and its definition is constantly changing.

Instead of trying to specify KSAs and relate them to performance at work, it would be more feasible—and arguably more useful for policy—to test whether practices intended to promote learning at work lead to better performance by individuals or groups. I have mentioned some of these practices: cross-training, job rotation, and skill-based pay. These are all intended to promote the transmission of knowledge and skill from those who have it to those who want or need it. In addition, some organizations have procedures designed to promote the discovery of new knowledge in the work process. Cross-cutting this distinction, it may also be useful to classify workplace education practices by whether they take place "on-line" in the actual work setting or "off-line" in a classroom or other instructional milieu. This yields a four-way classification, examples of which are as follows:

	Transmission	**Discovery**
Off-line	classroom instruction	quality circles
On-line	job rotation, cross-training, skill-based pay	procedures to elicit suggestions for continuous improvement

Adult education surveys could include questions about participation in these and other arrangements for workplace education. In particular, teacher surveys could measure the prevalence of these practices in their workplace, which is the school system. Further, the association between participation in such arrangements and the work performance of individuals or groups could be measured. If the study is longitudinal, it would also be possible to measure the correlation with performance in subsequent work settings for individuals who change jobs.

Such studies would begin to illuminate whether and how education in the workplace affects performance at work.

How Work Affects Education

Cappelli also correctly emphasizes the fact that most students hold paid jobs while in high school or college (see also Stern and Nakata 1991). Indeed, the 1994 School-to-Work Opportunities Act encourages schools to incorporate more "work-based learning" into the curriculum. One logical justification for this policy is the expectation that students will become more capable of learning at work as adults if they practice doing it while in school. A study in France, where detailed statistics are collected on adult learning at work, indicates that individuals whose initial schooling included some work-based learning do, in fact, participate more in continuing education at work (Romani and Werquin 1995).

However, as Cappelli points out, most research in the United States on the effects of students' employment has considered only the amount of time they spend at work, ignoring qualitative aspects of their work experience. A recent exception is a longitudinal study conducted at the National Center for Research in Vocational Education, which has discovered correlations between certain characteristics of students' work and their school performance, as well as with their wages a few years later (Stone et al. 1991; Stern et al. 1995, and Stern 1996 forthcoming). NCES could build on this study to incorporate questions about students' job characteristics into longitudinal surveys of students, both K–12 and adult.

Conclusion

Traditionally, the connection between education and productive activity has been considered to be primarily sequential. Now it appears to be increasingly synchronous. NCES is in a position to provide essential data for describing and understanding the consequences of this convergence.

References

Romani, C. and Werquin, P. 1995. "Alternating Training and the School-to-Work Transition: Programs, Assessment, Prospects." Paper presented to the CERI/NCAL Roundtable on School-to-Work Transition in OECD Countries, Paris, February 2–3.

Seavey, D. and Kazis, R. 1994. *Skills Assessment, Job Placement, and Training: What Can Be Learned from the Temporary Help/Staffing Industry?* Boston: Jobs for the Future, Inc.

Stern, D., Finkelstein, N., Stone, J.R. III, Latting, J., and Dornsife, C. 1995. *School to Work: Research on Programs in the United States*. Washington and London: Taylor and Francis, Falmer Press.

Stern, D., Finkelstein, N., Urquiola, M., and Cagampang, H. March 1996 forthcoming. "What Difference Does it Make If School and Work Are Connected? Evidence on Cooperative Education in the United States." *Economics of Education Review*.

Stern, D. and Nakata, Y. 1991. "Paid Employment Among U.S. College Students: Trends, Effects, and Possible Causes." *Journal of Higher Education* 62 (1): 25–43.

Stone, J. R. III, Stern, D., Hopkins, C., and McMillion, M. 1990. "Adolescents' Perception of Their Work: School Supervised and Non-School-Supervised." *Journal of Vocational Education Research* 15 (2): 31–53.

9

Using Administrative Records and New Developments in Technology

Administrative Record Opportunities in Education Survey Research

Fritz Scheuren

INTRODUCTION

This paper addresses possible administrative record opportunities in the education survey research work of the National Center for Education Statistics (NCES). Elementary, secondary, and postsecondary education are included. The time horizon is roughly the next decade—through 2005—but some discussion will be provided that extends beyond that period, primarily in connection with the Decennial Census of 2010.

Organizationally, the paper is divided up into seven sections—(1) this introduction with some background and other introductory materials; (2) a look at overall trends that have an impact on administrative record electronic access; (3) possible scenarios in education statistics; (4) survey investment opportunities arising out of those scenarios; (5) related analysis opportunities and barriers; (6) the privacy and security issues that must be faced; and (7) a conclusion with an overall summary and some recommendations. An attempt has been made to keep the prospective broad, drawing on themes that are emerging or have emerged in statistical uses of administrative records generally. Naturally, there is particular emphasis on current NCES surveys.

Motivation and Goals

There is a widespread sense that the U.S. education system needs major improvement and that one way to help achieve this is through better statistical information systems (e.g., U.S. Department of Education 1991). NCES has produced an extensive array of survey products and related publications to address the need to monitor progress (see, especially, U.S. Department of Education 1994a). Many of the surveys it employs are based in whole or in part on administrative record data; however, still greater use of administrative records may be possible and it is the purpose of this paper to explore that option.

Scope of Administrative Records Examined

Formal administrative records in elementary and secondary schools and local education agencies are of six types:

1) Pupil records (cumulative folders, transcripts, etc.);
2) Instructional service records (courses offered, textbooks used, etc.);
3) Personnel records (specific teaching assignments, certification level, college transcripts, etc.);
4) Financial records (accounting journals, payroll records, etc.);
5) Records required by other agencies (health records, W-2s, etc.); and
6) Policy records (special tabular analyses and reports, etc.).

This list comes from a 1985 report prepared for the Office of Educational Research and Improvement (Hall et al. 1985). That source goes on to note that some records are initially maintained in separate school files and then summarized and entered into central record systems. The detailed content and organization of these files can vary from one local education agency to another—causing massive reporting and summarization problems as the records are further processed for use at the state level or for other entities, including NCES.

The information contained in these reports forms the core of a state's information system on local education agencies and schools. Supplementary data collections, including student testing programs, also occur. In 1985, though, for most states, these had not been integrated into a comprehensive educational information system.

Frankly, it is unclear as to the extent to which the above description of problems with administrative data continues to be true. There is evidence that matters have improved, with at least some states becoming highly automated (U.S. Department of Education 1994b). The overall extent of the progress being made is, however, unknown. Certainly, there have been some highly successful prototypes, notably in Nevada (Nevada Department of Education 1994).

The administrative records used for postsecondary education generically are similar to those for elementary and secondary institutions, with, of course, some important additions—like data from the federal student aid program. The impression is that colleges and universities, at least the large ones, are much further along in building integrated, electronic administrative databases.

One obvious recommendation to make for the future is to consider routinely and systematically tracking progress on improvements in the record management practices of at least a sample of the 15,800 local school districts and 85,000 public schools. Knowing how automation is proceeding in postsecondary institutions also should be routinely monitored, with success stories shared as appropriate.

In the next section there is a general discussion of overall trends that might affect strategies with respect to administrative record opportunities in NCES surveys. This is intended to frame the specific options that will be covered later.

SPECULATIONS ABOUT TRENDS

Making credible any prediction about the future is obviously problematic—especially involving technology up to a decade or more from now (Rennie 1995); hence, the approach here will be to discuss "scenarios," rather than to actually make any flat assertions about what will or will not happen. To motivate the scenarios to be discussed, broad, mostly obvious trends are speculated about below. These have been divided up into trends in computing, costs and budgets, survey science, institutional change, and concerns about personal privacy. What has been highlighted is not necessarily what is most likely to happen; indeed, in some cases, items are mentioned mainly because, if they did happen, great changes would have to be made in the way NCES currently does business.

Computer Technology

Among the possible computer technology trends (e.g., Ligon 1996) that bear on administrative record opportunities in education survey research are:

- Low cost personal and organizational computing power continues to spread ever more widely.
- Advances in telecommunications make possible the movement of increasingly large masses of data. The National Information Infrastructure effort is a major reason for this (Office of Science and Technology Policy 1994).
- Both of these trends are supported by increasingly powerful commercially available software.

The computing changes here are not only important in themselves, but they have opened up to many a whole new way to imagine the future. This has made people receptive to still other innovations too.

Costs and Budgets

A binding force that could harness these computer trends in a way that would increase information use of administrative records is what is happening to costs and budgets:

- Budgets could shrink greatly in all parts of the federal government (or at least not continue to rise).
- Costs of administrative data capture, in well-designed systems, would shrink too, perhaps at a rate faster than budget cuts.
- Costs of survey taking, on the other hand, have already dropped and would continue to do so, but at a slower rate (e.g., Nicholls 1988).

One implication of these observations is that administrative records will be much more available in an electronic medium than at present and at a lower and lower cost, relative to surveys. This possibility is a central motivation behind any expansion of administrative record opportunities in education survey research.

Survey Science

The revolution that began just 100 years ago (e.g., Bellhouse 1988), with the advent of representative sampling, shows no signs of being overturned; nonetheless, the role of surveys and censuses could be modified greatly in the next decade (e.g., Scheuren 1995). Given what has already been said about structural changes in computing and costs, it seems possible that:

- Both novel and traditional uses made of administrative records will increase in lieu of surveys or in hybrid combinations. The 1996 National Postsecondary Student Aid Survey (NPSAS) is a clear example of what can be done in building a successful hybrid (see the Appendix).
- Surveys may play a "Rosetta Stone" role—to adjust administrative data and to help interpret such data, rather than being relied on directly to make estimates.
- Microsimulation and other modeling (e.g., National Academy of Science 1992) based on administrative data, statistically matched perhaps to outside sources, will increase—along with other prediction/projection techniques—because large-scale direct survey estimates might be affordable only at increasingly infrequent intervals.

Randomization-based survey estimates will continue to be the "Gold Standard" against which other methods of creating information will be calibrated. Budgets, though, will not permit the sample sizes of today and, as this paper argues, cheap administrative substitutes should be sought from which to make generalizations, especially for small domains and small areas (e.g., Boruch and Terhanian 1996; Schaible 1996).

Institutional Change

"Third-wave" ideas about how to organize and run educational and other large institutions may be widely tried (Toffler and Toffler 1994). This may span the gamut from an even more serious look at Japanese quality innovations (e.g., Mulrow and Scheuren 1995) to the breakup of public schools, as we now know them. If changes of this magnitude (e.g., Newmann and Wehlage 1995) get going during the coming decade, they can be expected to materially affect the incentives for providing access to administrative records. "Charter schools," say, could have the same costs and budget pressures as the elementary and secondary public institutions they replace. Universities may be the most affected, since their costs have risen the most steeply and since they may be altered the most (Noam 1995).

It is possible, however, that institutional changes could speed up rather than impede the other innovations envisioned above. Even so, "the breakup of the old order," should it occur, may cause real stresses in comparability of and access to administrative records across the nation's schools and colleges. Much the same comparability problem exists now though, and there are several public jurisdictions (parts of Maryland and New York, say) where cooperation already seems tenuous at best—so what we may be dealing with here is more a matter of degree, rather than a major difference from the current situation (e.g., Salvucci et al. 1995).

Personal Privacy

Most of the trends mentioned above, arguably, could aid in increasing access by researchers to administrative records—at least not harming such an outcome greatly. Privacy and data security concerns, though, could slow down or permanently limit the growth in statistical uses of administrative records.

In some ways, NCES could not be better prepared to deal with such concerns. For example, the making and keeping of confidentiality pledges are nothing new at NCES; indeed, the Center has been an innovator in this area. Still the political debate coming could lead to legislation; predicting that law's effect on research will not be attempted here. Ways to mitigate any tradeoff between information needs and privacy exist though, and will need to be dealt with. For more on this, see U.S. Department of Education (1994b), where there is an extensive discussion of the early thinking of some of the state data stewards responsible for physical security and the protection of privacy. Use of the social security number, for example, is apparently already a sore point. It has been dropped from Virginia's student files. Moreover, access to any form of identifiable data, outside the local education authority, may be quite limited everywhere.

TWO SCENARIOS

Two scenarios are set out below. Each highlights what could be big changes from the current situation at NCES, relative to administrative records and their use with surveys. These scenarios are labeled, "Good" and "OK":

- "Good" is perhaps what one might want to happen.
- "OK" is a world that is livable but not desirable.

A "Poor" scenario was also looked at, but was so gloomy that it did not warrant writing about in detail.

The scenarios are both made up of a mix of the trends mentioned already, with a few natural extensions. As will be seen, there are common elements. Obviously, to the extent that the alternative futures set out here are credible, some of the commonalities noted may lead to anticipatory actions or investments on the part of the Center.

"Good" Scenario

Driven by concerns about international competitiveness (e.g., U.S. Department of Education 1994c) and the need to enhance the delivery of educational opportunities, a strong cooperative spirit continues in the education community, even as the current institutional structures undergo change. Some elements of this "good" scenario are:

- Smaller survey budgets almost certainly will be in store for NCES, but cuts will be modest relative to cuts elsewhere in the federal government.
- Despite tight budgets, NCES' role as an information coordinator and catalyst will be desired and clearly recognized, in both the public and private sectors of education. An example of such support might be the statement by the Council of Chief State School Officers, namely "We strongly urge that . . . NCES be a true statistical center that assumes the major responsibility for coordination of the collection, assembly, analysis and dissemination for that sector of society under its purview, namely education." (Hall et al. 1985).
- Further, the changes envisioned should be gradual enough to allow NCES' electronic interchange efforts to link virtually the entire educational system into a common network. The National Research and Education Network (NREN) will establish a gigabyte communications infrastructure to enhance the ability of U.S. researchers and educators to perform collaborative research and education activities, regardless of their physical location or local computational and information resources. This infrastructure will be an extension of the Internet, and will serve as a catalyst for the development of the high speed communications and information systems needed for the National Information Infrastructure (Office of Science and Technology Policy 1994).
- This network could give ready access to identifiable electronically available administrative data at the school and maybe even at the student level (albeit this last is problematic, as noted earlier).
- Traditional, mainly paper school records would become increasingly automated, allowing for the education network envisioned to supply administrative data rapidly and cheaply for statistical uses.
- A flexible survey system, evolved from current NCES efforts, will make it possible to interpret these administrative data and to augment them when necessary.
- Samples are likely to be smaller in size than currently, but regular, with all the economies gained by continuous production and refinement. This may be a hard thing to sell, but a careful look at the time series versus cross-section tradeoffs (e.g., Ghosh et al. 1995) might make the case—especially if some of the administrative record proposals in the next section turn out to have value.
- Each "node" in the education network (state education office, school, or school district) will be able to create its own custom products; hence, the pressures of competition will work to keep the system innovative and cost effective. This third-wave approach brings each organizational element into the system, in some sense, as an equal.
- Standardization of administrative records will only be partial and full standardization may not even be seen as desirable by participants. That lack of full standardization obviously could be a major cost barrier, unless there is a change in underlying thinking about what the data mean. Again, this is a third-wave notion, but this time already well accepted in accounting circles where each corporation—read school/district/state here

(?)—can, within generally accepted accounting principles (GAAP), decide how to keep their books. Once established and approved, of course, the entity must continue in the same way. Why can't this work in education?

- Some elements of the educational system may be unwilling or unable to cooperate in sharing administrative data and so provision must be made for these. Groups that will require special treatment might be children being "home-schooled" (an already large movement that is likely to grow even larger). Some institutions of higher education (say, "Ivy League" schools) may also not want to be involved for other reasons. This is occasionally a problem already.
- Privacy issues will need to be carefully addressed; nonetheless, they should not be a major barrier to research uses of educational administrative records. Physical security and monitoring systems for administrative electronic data used in research will be a major cost of maintaining trust in the network being envisioned. Training to enhance "Privacy Literacy" among researchers will also be needed—again at no small cost.
- Use of administrative records from other systems should also be possible, including tax data (Forms 941, W-2s—maybe even 1040s), but access will necessarily be more limited—maybe only on a sample basis and with special consent arrangements. Partnering with the Bureau of Labor Statistics to use unemployment insurance records might also be possible; at least it should be explored.

Two last comments before going on to describe another scenario: (1) NCES may not be the only major information supplier in this networked world. Privatization is a distinct possibility. A lot will depend on how well the Center adapts to the changes coming. (2) It seems likely, though, that NCES could make "leading" contributions to developing the needed education information systems for this world. ("Leading" and "running" are not the same. It might be very undesirable for NCES to try to dominate in this networked world. To accomplish its mission, all it needs to be is a major player.)

"OK" Scenario

Again, as above, there is seen to be a compelling need by all to cooperate in achieving national education goals. More barriers to change exist, though, in this scenario, and a "limited success" is all that occurs in the coming decade. Some of the elements in this only "OK" world are:

- Declining survey budgets occur for NCES; the cuts, though, will be about the same as the average of cuts in statistical programs elsewhere in the federal government.
- Even so, a clear role for NCES as an information coordinator continues to be widely accepted, in both the public and private sectors of education. Resources to act as a catalyst in broadening administrative record research uses are, however, necessarily limited.
- Plausibly, the budget changes envisioned may not be gradual enough to allow NCES' electronic interchange efforts to link virtually the entire educational system into a

common network. Still, most of the system could be networked anyway, but closer to the end rather than the beginning of the coming decade.

- This network would provide, as above, ready access to at least limited identifiable administrative data at the school but probably not at the student level.
- A flexible survey system, evolved from current NCES efforts, will make it possible to interpret these administrative data; for cost reasons, however, augmentation by direct surveying could be much less frequent than at present.
- Sample sizes are likely to be smaller as well, with few of the economies gained through continuous production and refinement.
- Some school or school district "nodes" will be able to create their own custom products, but this will not be an information-rich world—in many ways, information services may be about at the level they are today.
- Standardization of administrative records will be quite limited; however, developing and maintaining a metadata system, for at least the important concepts, should be attainable.
- Certain groups, like "home-schooled" children, despite their growing importance, will have to be ruled out of scope for most purposes.
- Privacy issues will need to be carefully addressed but still are not expected to be a major barrier to most research uses of available educational administrative records. Physical security and monitoring systems for administrative electronic data may be a concern in the network being envisioned, because only a "bare-bones approach" may be affordable. Training to enhance "Privacy Literacy" among researchers will have to be modest, exposing the system to a greater risk of a potential loss of trust on the part of the public.
- Use of administrative records from other systems could be very limited because of privacy and resource restrictions.

In summary, for this so-called "OK" scenario, NCES will at best be where it is today, except that inevitably budget cuts will have limited its information products at least somewhat. It is hard to imagine NCES leading, let alone running, the nation's education information systems in this world.

NCES Investment Opportunities

In the next section, we return to the overall trends mentioned earlier and suggest in broad terms what investments NCES might consider to increase the chance that the survey opportunities available in administrative records are enhanced—that is, that the "Good" scenario wins out over the only "OK" one.

SURVEY INVESTMENT OPPORTUNITIES

Administrative records play multiple roles in NCES surveys. Existing practice seems, therefore, to be a natural starting point for looking at further opportunities. Each of the major ways, current and proposed, where administrative records could be employed is discussed below, one at a time.

- *Administrative tabulations as a source of general information* are seemingly ubiquitous already. New opportunities here, if there are any, would lie in speeding up the availability of this information and potentially customizing it. On-line access is already fully in place for regularly prepared "ED TABS" summaries—e.g., as described in U.S. Department of Education (1994a); but see also what is being done elsewhere (Federal Committee on Statistical Methodology 1995).

- *Administrative data as a sampling frame* is very common too—at the school, teacher, or student level—e.g., in the public school components of the School and Staffing Survey (see McMillen, Kasprzyk, and Planchon 1993). Many opportunities exist, though, in this area. This is so especially if more data become electronically available on these frames, and quality improvements continue (e.g., Peng, Gruber, Smith, and Jabine 1993). Also, the time gap between the frame items and their potential survey use should be shortened; right now this can be up to 2 years or more.

- *Augmenting survey data with administrative items* during or after fieldwork is done in some NCES survey settings (e.g., NPSAS). Again, the opportunities for greater use of administrative records lie mainly in widening access to timely, electronically available data of high quality (U.S. Department of Education 1994c). Significant survey cost savings are obviously possible when comparable administrative data can be used, instead of obtaining the item by a direct survey method. The biennial NCES High School Transcript Study might be a place to begin to shift from the abstraction of data from paper records to direct electronic access. Differences in formats from state to state and even within states could be a major barrier, but a pilot might still be worth considering.

- *Editing survey data by comparing it to administrative items* is quite common in the establishment surveys of other agencies, such as Statistics Canada, the Bureau of Labor Statistics (BLS), or the Census Bureau. This use in NCES surveys seems to be infrequent at present, perhaps due to the timing and content of the administrative records that the Center has ready electronic access to. The Center has already sponsored studies (McMillen et al. 1993; Peng et al. 1993) which point to the possible benefits here, and pilot efforts to operationalize administrative data for editing survey variables might be among the steps to consider next.

- *Imputing for missing survey data using administrative records* is another common occurrence in establishment surveys at BLS and Census. Sometimes the administrative data are simply substituted directly; sometimes elaborate models are employed. It seems likely that both item and unit nonresponse (and perhaps coverage) adjustments could be improved if administrative data were employed. To test this idea out, NCES might want to conduct a pilot effort, say, with SASS and the Common Core of Data (CCD). This seems especially appropriate since so much analysis has been done recently with CCD and SASS. Of particular note is that CCD is available every year. One year's CCD can be used, thus, as a frame while a later year can be used to edit the survey and impute for missing or erroneous entries.

- *Expanding the uses of administrative records that come from outside the education community* may be an important place to invest more. Privacy and security issues obviously are key here. Enhancing this option, through improvements in record linkage techniques, could even be a priority—especially for higher education, where IRS income data might become available because of the student loan program (National Academy of Science 1993). The ubiquitous social security number (SSN) seems the practical choice for student and teacher linkages, provided the SSN is backed up by confirmatory variables (such as names, addresses, and birth dates). School linkages to, say, Form 941 data or to unemployment records, should these be possible, would pose still other challenges.

  However, there may be a problem with this obvious approach, as already noted earlier, because of privacy considerations. Additionally, there are technical issues in the record linkage itself, especially without an exact identifier (e.g., Alvey and Kilss 1985; Newcombe 1988; Newcombe, Fair, and Lalonde 1992; Belin and Rubin 1993; Winkler 1995; Winkler and Scheuren 1995).

  Minor housekeeping improvements between NCES survey systems (and within such systems over time) might be looked at to see how broadly conformable they are to linkage, either using exact or statistical matching techniques (U.S. Federal Committee on Statistical Methodology 1980). The routine addition (or use) of check digits for all "unique number identifiers"—including for schools—is a suggestion for cases where they are not already on the survey or administrative records being employed by the Center. Achieving common formats for items that might be used to do statistical matching across administrative and survey systems also seems to be another option to look at.

- *Weighted survey estimates, obtained by poststratification to administrative totals*, might allow NCES to reduce current sample sizes and save money, without increasing the variance of major statistics. This could be done simply by employing conventional ratio estimation, using administrative data on the frame for both sampled and nonsampled cases. See Kaufman, Li, and Scheuren (1995) for more powerful and general methods too. Conceivably, even frame data that is a year or two old might be worth experimenting with. Better, more timely administrative data, of course, could lead to even better results.

- *Longitudinal surveys can particularly benefit from available related administrative data.* Administrative data can be used to help track cases (e.g., address changes) between interviews. Changes in administrative items may be predictive of similar changes in survey variables—among both respondents and nonrespondents. Clearly, editing and imputing longitudinal survey variables are greatly strengthened, if longitudinal administrative data have been linked. Times between successive interviews may be stretched out too, resulting in cost savings. Longer gaps between interviews, of course, would work only if the administrative data are near substitutes in the nonsurvey period. Staggered panels that have some direct data collection every year but at wider intervals might be worth experimenting with, too, because of their potentially flexible, low cost nature.

- *"Mass imputation" of sample survey data to a complete population file* has been shown to work in some Canadian applications (e.g., Whitridge, Bureau, and Kovar 1990) and has advantages for NCES over simply weighting up administratively matched survey data. Mass imputation is a technique that assigns a survey case to one or more nonsampled cases in the population, using the overlapping data in some form of statistical matching. Each unit in the population is imputed a survey case. When efficiently done, the costs of mass imputing are only moderately larger than weighting. Recent work at the National Center of Health Statistics (NCHS) by Schafer and others in a Bayesian context provides an illustration of some of this method's real strengths—albeit for imputing for nonresponse (Schafer et al. 1993; Schafer 1991). Cheap computing is needed at the analysis stage because the whole population has to be processed. Given this last observation, it is not surprising that the Canadians, at only 1/10th the size of the U.S., were pioneers in this method. Nonetheless, the time is coming when the old computing cost barriers will be a thing of the past (even in government).

- *Mass imputation for small area estimates is also attractive* in an environment rich in detailed administrative data. Cross-section administrative data, like the Common Core of Data (CCD) for public schools, would be an ideal file to employ in experimental efforts to make small area estimates. To check this approach, a sample of areas—say, local school districts—would need to be selected. Direct survey observations in these selected areas would then be augmented sufficiently to test the idea. Obviously, for variables not closely related to those on the CCD not much should be expected—illustrating yet again the importance of expanding administrative items on NCES frames. The work NCES does with administrative records for small areas should, of course, not be confined to mass imputation, albeit mass imputation seems the most promising of the alternatives at this point (for more on small area estimation, see National Academy of Sciences 1992; Purcell and Kish 1979; Malec and Sedransk 1995; Schaible 1996).

- *Making survey time series estimates employing administrative data* is a natural extension of the methods being discussed. Initially, suppose that mass imputation techniques continue to be used. The step (leap) is from mass imputation (to cross-section administrative records) for small area estimates to doing mass imputation (to longitudinal administrative records) for time series estimates. Both start out with direct sample observations. In small area estimation a model is developed which predicts what the nonsampled cases would have reported in the survey for each element in the population in each area of the country. It is just one further, albeit big, step to predict what would have been reported by nonselected *and selected* cases, if the survey had been done again in, say, a different year. Obviously, changes in administrative data would be additional factors to consider in the imputation; that is, once an initial small area estimate had been made through imputation, it could be a starting point for small area *and time series* imputations for the next year. Time series estimation is an even older and deeper field statistically than is small area estimation; hence, other methods besides mass imputation ought certainly to be tried. Whatever is finally done, the need to check on the estimates by direct survey measurement exists here too and could be a source of improvement ideas as well as helping to interpret the results.

- Administrative records, as indicated above, can be used by NCES in both novel and traditional ways. Some of these NCES has already been developing. In each example, though, the starting point was a survey. *What if the starting points were the administrative records themselves*, as is the case for most samples in some other agencies (e.g., the IRS)? In this later world, the main emphasis shifts to processing the administrative data and to using them directly for inference. Surveys could play a "Rosetta Stone" role—to adjust administrative data and to help interpret such data, rather than being relied on directly to make estimates.

At the outset of this paper, it was conjectured that randomization-based survey estimates would continue to be the "Gold Standard" against which other methods of creating information are calibrated. In this context, it was also said that future NCES budgets might not permit the sample sizes of today. Cheap partial (or complete) administrative record data might be appropriate substitutes, especially for small domains and small areas. As we have seen already, there are many ways for NCES to continue to take steps (big or little) in this direction. The implications of this "brave new world" for analysis, and analysts, will be covered next.

ANALYSIS OPPORTUNITIES AND BARRIERS

The previous section began with the existing ways that administrative records now support NCES surveys. Some ideas were also given on possibly strengthening these conventional methods. Gradually, though, the ideas for change moved more and more away from pure randomization-based survey inference; progressively, they were replaced by modeling ideas of various sorts (e.g., Särndal, Swensson, and Wretman 1991; Smith 1994).

Even supposing all of these ideas were sensible—and some of them undoubtedly will not work out—what would the benefits be? Is all this change worth the trouble? Is it possible that in order to save on data capture costs, other costs are being incurred that might be very large? Are costs being shifted from data producers to data users? Well, if a one-word answer were to be given, it would have to be "Yes"—at least some of the time. The old saw is also partly true, "We are trading the devil we know for the devil we don't." Unquestionably, one set of hard problems is being replaced by another.

Just look at the "Rosetta Stone" comment made above. While admittedly the most extreme of the options, this approach would be enormously challenging for educational researchers. In this world, surveys might be a much smaller part of the database, with many of the files being almost purely administrative. In such cases, survey vehicles would be used only to lightly monitor and interpret ongoing administrative data and to help explore new areas where administrative data did not yet exist—perhaps in an experimental setting or as part of an observational study of a new educational alternative.

This nearly completely administrative data world is not likely to happen soon—and for some information requirements, like opinion data, probably never. First, a much richer, fully networked, administrative data set is needed. Second, the eleven other options listed in the last section ought to be considered and maybe tried too—moving from those that are only modest extensions of what is now being done, to those requiring bigger and bigger changes on the part of both data producers and data users. Some additional steps are also recommended. Three of these are discussed in the subsections which follow.

Shifting the Emphasis From Data to Information

Understanding better the ways that current NCES data are turned into information by the Center itself, or through outside users, is an essential and obvious step. Data are products that, to be useful, must be "enlivened" by users. It is only through a positive synergy among data, data producers, and data users that information arises. Metadata systems are one of the best ways of making this synergy more systematic and more often fruitful. Strengthening Center efforts should be considered here, if only as a way of better tracking changes over which the Center has no control. NCES already does an outstanding job in running user training workshops and bringing interested individuals fully in contact with the data that the Center produces. Although already good, better file documentation is needed. Benchmarking studies on the metadata systems that other agencies (particularly administrative ones) are building might be a useful way to get potentially workable improvement ideas.

Further shifting of Center emphasis to providing information services rather than tabulations and data products cannot be stressed enough, as a way of preparing Center staff for the future discussed in this paper. Said another way: It is essential to look at the work being done from the customer end—realizing that all customers cannot be satisfied, even though that still should be the goal. Typically, data systems are very sluggish and change slowly. Information needs, on the other hand, move much more rapidly. A Center goal might be to develop *information systems* that are rapid, even though the *data systems* to which they are anchored may not be.

If, as seems likely, there will be more work for users to do as a result of the changes discussed in this paper, then one simple strategy is to *find more users to do it*. This admittedly "Tom Sawyer" approach is only a partial answer but it could help. Users have increasingly more powerful computing, possibly better than what NCES has, so big files and complex data structures may be seen as a welcome challenge to some—especially if there are more data overall and the data can be made more timely. In short, a marketing strategy might be warranted, and perhaps in market segments that are outside the traditional research community. With the proper privacy safeguards in place, these segments might include school administrators and other operating personnel (teachers and students?) at all levels of the national education system, who might want to compare themselves to those in similar circumstances. This expansion of users could go naturally, hand in hand, with a broadened access to secure administrative data for research purposes.

Getting the Distributions "Right"

Shifting to methods which emphasize more the need to get the "inference right," rather than just getting the "data right," seems essential. What does this mean? At present, most statistical agencies around the world spend a sizable fraction of their resources in collecting data and cleaning up inconsistencies in them—in short, on getting the "right data" (e.g., U.S. Federal Committee on Statistical Methodology 1990). Because these agencies are invariably peopled mainly by nonstatisticians, the idea that the data come from some underlying distribution with inherent uncertainties in it can get lost.

Technically, what is needed is to understand these distributions—to get them "roughly right," as Tukey has said. The data are thus only a means to an end, *not the end!* In a way, this is the same point made earlier, when information systems were being discussed. It is upon these distributions (Rubin 1990)—whether parametric or nonparametric, formal or informal—that inferences get made. The underlying causal mechanisms (or distributions) that generate the data observed are models that may, in the eye of the observer, be suggested by data or which can be fit to data. Distributions, thus, are a construct of the questioning observer. Obviously, the notion of distributions, then, unlike data, gives the user the central role.

"Selling" the user on data obtained from administrative records must be done for such records to be the basis for the creative leaps that research must make when new knowledge is borne. How might this be done? Assume two variables, one administrative and one survey-based, are compared and a scatterplot constructed that shows a strong relationship. Should the two variables be highly related, then arguably the same inference might be made from either one of them. Even so, the administrative variable might not be defined in quite the way that the researcher would like. On the other hand, the survey variable, while definitionally more suitable, could be costly to get; moreover, the survey variable would still be subject to sampling and measurement errors that could impair its use for inference. It truly is a question of deciding between the devil you know and the one you don't. Only experience will tell which devil is easier to live with. In any case, increasing reliance on administrative data may require experiments of the sort implied by this discussion. There is a lot at stake here. Put provocatively, should NCES invest in methods that may not even be based on exactly the "right data" but that could, most of the time, yield the "right inference" anyway? If the answer is "Yes," how might this be done? Beyond the answer, "it depends," not much of general value can be said here. Each such decision will need to be looked at individually.

Still, there is at least one comment worth making. With greatly expanded access to administrative data, the resources to do the careful (over)editing (Granquist and Kovar 1995) now characteristic of most survey systems would literally be impossible to find. Choosing new summary statistics that are robust against data problems is one obvious suggestion: medians instead of means, interquartile ranges in lieu of variances, graphical displays rather than tables of totals; all could allow users to see a distribution's shape in the presence of messy data. These or better methods make sense in the presence of administrative data of the scope envisioned. If the data suppliers are also data users (see "Shifting the Emphasis From Data to Information," above), then some of the Japanese quality improvement ideas might take stronger hold, leading to less back-end editing but without any sacrifice in "inference quality."

More Emphasis on Measuring Uncertainty

Strengthening the Center's efforts to measure sampling and other forms of uncertainty seems crucial too. At this point NCES has made great strides in building survey information systems that allow the user to measure sampling error. This is no small feat, given the complexity of the data collection. Much more will be needed, though, for the administrative data environment envisioned.

Some of the issues that will have to be addressed already exist today. For example, quantifying uncertainty in the presence of imputed data is an area of controversy at this point in surveys (e.g., Rubin 1996; Fay 1996). Mass imputation methods are not immune from criticism either (Rubin 1990). In a mass imputation world, of course, the administrative data would not be subject to sampling error. As far as the survey data go, they could have variances calculable, via methods that adjusted for the implicit poststratification that the imputation should generate (e.g., Wong and Ho 1991). How to estimate mean square errors for the joint distribution of survey and administrative data is an area that has been studied but seems to need more (basic?) research.

Among the tools being employed by NCES at present, resampling ideas, such as bootstrapping techniques, could be the best place to make further investments in estimating sampling variances (for more on bootstrapping in an NCES context, see, for example, Kaufman 1995). Gibbs sampling tools could help, too, if more general measures of uncertainty were desired.

Winners and Losers

In this section, three analysis issues have been briefly discussed in the context of a possible large-scale expansion of administrative record use—with or occasionally in lieu of NCES surveys. The topics covered were illustrative and not exhaustive:

- To focus more on the information end, rather than the data end of the Center's work;
- To reallocate resources away from data cleaning[1] and toward better ways to see underlying distributions; and, finally, and very briefly
- To look hard at techniques to measure uncertainty that work during the period when the transitions envisioned will be taking place.

Clearly, if and when the most radical of these administrative record changes came about, there would be major consequences for education researchers. Since the time span is so long—10 years or more—and given that small experimental intermediate steps are possible, adjustment problems seem manageable. This is not to say that adjustment will be easy; in some places they can be predicted to be hard indeed.

PRIVACY, CONFIDENTIALITY, AND DATA SECURITY

Privacy, confidentiality, and data security issues have been given considerable attention in many forums in recent years. The range of treatments is a wide one, spanning the 1993 book, *Private Lives and Public Policies*, which focused on research data and was intended for specialists, to the very recent book, *The Right to Privacy*, which, while also a considerable scholarly accomplishment, is intended for a more general audience (Duncan, Jabine, and de Wolf 1993; Alderman and Kennedy 1995).

Tore Dalenius has provided a good review of privacy, confidentiality, and security goals in statistical settings. His work may afford a point of departure here (e.g., Dalenius 1988; see also Boruch and Cecil 1979). In common speech, the words privacy, confidentiality, and security partially overlap in usage and often have meanings that depend greatly on context. Each can also have an emotional content which makes precise definitions difficult, even contentious. For example, Dalenius quotes Westin (1967) about privacy: "Few values so fundamental to society as privacy have been left so undefined in social theory or have been the subject of such vague and confused writing by social scientists."

A good start on giving meaning to the word "privacy," or "information privacy" (our context here), might be the definition first articulated by Justice Brandeis as the "right to be left alone . . . the most comprehensive of rights and the right most valued by civilized man" (Olmstead 1928). With books like *Private Lives and Public Policies*, it appears we may finally be making serious progress in operationalizing the Brandeis definition of "privacy rights"—at least as they relate to statistical information. Much remains to be done though. The practices of data stewards in education (U.S. Department of Education 1994b), and elsewhere (Jabine 1993) vary widely. Public opinion research shows a range of concerns, too, depending on the context in which questions about privacy are asked (Scheuren 1985, 1995; Blair 1995; Presser and Singer 1995). Information on informed consent exists too but is dated (Singer 1993).

The National Center for Education Statistics has been the pathbreaker in giving controlled access to its survey files for qualified researchers (e.g., Wright and Ahmed 1990). The Center needs to continue taking the same kind of leadership position with regard to assuring wide educational research access to administrative records as it has with surveys (e.g., U.S. Department of Education 1995).

The final outcome here, though, is quite uncertain, since each state may legislate separately on the kind of electronic access that will be permitted for statistical purposes. Identifiable school level data are already extracted (in CCD). Having identifiable student level administrative data at NCES would be desirable, for example, for many surveys too. Overall data security issues deserve NCES attention, particularly as electronic administrative data become more and more widely available. In this regard, the recommendations[2] in the report *Educational Data Confidentiality* (U.S. Department of Education 1994b) are worth quoting at length:

> There appears to be a need to inform those who work with electronic data and citizens as well as taxpayers of laws, regulations, and procedures that schools, states, and regional agencies adhere to in collecting, using, and protecting data confidentiality. Such information should be widely available, readable, and easily

understood. It should summarize current federal and state assurances of privacy and limits on data access and use, and be accessible to the public through government agencies at local, state, and federal levels. These central findings are suggested:

- Standards, procedures, and recommendations are available from other agencies, and from states that have established workable procedures, but there is relatively limited cross-agency or cross-state exchange, and wider dissemination of models would advance the security of new systems.
- States and other data agencies should be encouraged to inform agency personnel who work with personal record information—including student records, personnel records, and family demographic information—what regulatory restrictions limit access and use and encourage staff persons to make an effort to keep members of the public well informed of these rules, assurances, and routine protections of privacy.
- States, districts, and other data agencies need more routine procedures for publicizing widely across agencies and among taxpayers and citizens the confidentiality protections they have in place.

Some areas where emerging issues may need monitoring are mentioned below. It should be noted, that while this list has many challenges, it is by no means exhaustive. These are:

- How to manage the physical data security for this new information network, so that the system is fully "auditable"—i.e., access records are kept of what was looked at, by whom, what changes are permitted and get made (for more here, see Brannigan and Beier 1995).
- How to assure that proper notification and consent procedures are followed so that individual human rights are respected (e.g., Singer, Shapiro, and Jacobs 1995; Scheuren 1985; Scheuren 1995). Continuing experiments seem the wisest course here and might be worthy of consideration by NCES.
- How to adjust for cases where consent is denied to administrative records by NCES survey respondents. This is a very tough problem if the refusals are at all sizable, which does not seem likely at this point. Basically what seems needed is to institute statistical work on group matching or other techniques that would lessen the tradeoff between the competing values of furthering scientific research *and* safeguarding personal privacy (e.g., Spruill and Gastwirth 1982; Gastwirth and Johnson 1994).
- How to track public opinion on the education research uses being made of the linked data network being built. The series of Harris-Equifax surveys are one source here, albeit imperfect (Harris et al. 1993). The Harris-Equifax surveys have important limitations (Blair 1995) on their interpretability; nonetheless, their main conclusions are in essential agreement with other research on privacy concerns. Roughly, almost no matter how you ask the question, there are always about one sixth to one fifth of the population who oppose electronic record linkages on privacy grounds. Conversely, again almost no matter how you ask the question, about the same fraction will favor

"beneficial sounding" linkages on efficiency grounds. The two thirds or so in the middle will differ in their opinions depending on the specifics. See also Presser and Singer (1995).

- How to protect research data from nonresearch uses, especially by governmental entities. See, especially, Chapter 1, *Private Lives and Public Policies* (Duncan, Jabine, and de Wolf 1993).
- How to reduce inadvertent reidentification risks, especially those that arise through school level linkages with student data. This is the same problem, in some ways, that exists with the Social Security Administration's Continuous Work History Sample (CWHS). See, for example, Jabine and Scheuren (1985).

Clearly, privacy and related confidentiality and security issues must continue to be faced as administrative data become increasingly available electronically. The uncertainties about the future seem greater here than elsewhere but the Center has done a lot already and seems poised to do more.

SUMMARY AND RECOMMENDATIONS

In this paper, there has been a broad discussion of opportunities for making more effective use of administrative records in surveys of elementary, secondary, and postsecondary education. Here it may be appropriate to group what has been said concerning:

- *How is the availability of data in administrative records likely to change during the next decade and how will these changes influence opportunities for NCES data collection and analysis?* The first three sections of this paper cover topics in this area. Predicting the future is so difficult that two scenarios were used and formed the basis of discussion. Many specific suggestions were made in passing. The underlying premise, though, as far as basic research is concerned, is that the Center probably cannot afford to make major financial investments. Staff investments are needed, nonetheless, in monitoring the changes coming and mining them for ideas to try in ongoing Center efforts. The Center's role as a technology transfer catalyst is where investments should be made, if possible, in bringing the good ideas on-line faster. It might be necessary to help bring cheaper administrative data capture and electronic transfer technologies to schools so that they can lower or at least contain these "back-office" costs—much as banks and insurance companies have begun to do. What, for example, can the Center do to help create and test cheap scannable forms for some routine transactions and bankcard-like direct electronic access for others?
- *What are the opportunities for better integrating surveys of individuals (for example, students, teachers, administrators, or parents) with existing administrative records to improve the quality and utility of NCES surveys?* Here the Center can and needs to do the most. Imbedded experiments with new methods of design, data capture, estimation, and analysis should be a growing part of NCES survey efforts. A whole range of these was discussed under "Survey Investment Opportunities" and "Analysis Opportunities and Barriers," above. In particular, work involving CCD is a natural place to make a

concentrated effort (Holt and Scanlon 1994), but following up on the design ideas in the 1996 NPSAS makes a lot of sense too (see Appendix). Center tradition and recent research supports such growth. There is, however, the usual problem with all surveys of being conservative about change, once a survey has begun to operate (Dillman 1994; Groves 1995). Looking at survey contract vehicles and staff incentives will be crucial to overcome the natural risk adverse behavior that is likely to exist.

- *What are the main issues or barriers surrounding access to or better use of administrative records, and how might these be addressed?* Throughout the paper, but especially in the "Privacy, Confidentiality, and Data Security" section above, there were places where the many barriers to change were dealt with. Some of these can be overcome by gaining new knowledge, e.g., by more methods research on, say, CCD—perhaps an experiment to directly access school administrative records, rather than continue to transcribe them as at present.[3] For many issues, a wait and see approach may be the only strategy possible. Especially for changes in institutional arrangements, the Center probably has no role, except to react to events. There are still activities to be considered, however. For example, developing generalized capabilities to react is one option here. In the context of administrative record access, for example, in the privacy area continuing to work toward a fully secure network environment for research, auditable by each school and even each student, could go a long way to overcome potential concerns.

Still another activity that might be emphasized, in the Center's applied research, is the private school segment of elementary and secondary education. An extra effort in this area would warrant consideration, depending on what seemed the likely speed of movements to change the "Old Order," such as the creation of Charter schools. What about experimenting with partial Internet-available (encrypted?) administrative record alternatives to the Private School Survey?

AN AFTERWORD

Of course, even if the Center does not try to speed up beneficial change, change is inevitable and, hence, the Center will have to deal with imbedded experiments involving all sorts of changes, including to administrative records. The question is what role will NCES take in their design or even whether they get designed or just happen. Crucial, too, is how will the Center protect its surveys when these experiments go wrong, as occasionally might occur, no matter how well they are designed.

One of the most encouraging things is that those who welcome the future changes coming will not be alone. Virtually all large organizations are moving in the same direction (Nanopoulos 1995), even statistical ones (e.g., Keller 1995). The positive synergy from the massiveness of what is happening should sweep up those organizations, like NCES, who want to change and move them much farther than their individual efforts alone would make possible.

The way these outside changes play out within the educational community may deserve the most staff attention. As noted already, especially important from an administrative record perspective is speeding up the automation and networking of administrative records, since without broadened access and drastic cost cuts, such records will largely remain on paper and, hence, hard to obtain for survey research.

One final point, however optimistic one may be about the (distant?) future, there is a long way to go. The *Wall Street Journal*, in a special section on school (mainly computer) technology, dated November 13, 1995, made this point extremely well (*Wall Street Journal* 1995; see also *Science* 1995). The Center has, though, clearly made a good start. It is hoped this paper will help too.

NOTES

1. This observation may seem in conflict with one of the recommended additional uses of administrative records (i.e., for editing) discussed in the previous section. The issue is not to stop editing, but to stop overediting, a point made in the 1995 Granquist-Kovar paper cited earlier.

2. Also of interest are the views expressed in the report of the Privacy Working Group on the U.S. National Information Infrastructure (1995).

3. At present, the CCD is not processed as a longitudinal file. Longitudinal (transaction-based) processing is another important improvement to consider, especially anticipating the day when administrative data are put on the CCD directly without a separate extraction step.

REFERENCES

Adelman, C. 1995. *The New College Course Map and Transcript Files: Changes in Course-Taking and Achievement, 1972–1993* (based on the Postsecondary Records from Two National Longitudinal Studies). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, National Institute on Postsecondary Education, Libraries, and Lifelong Learning.

Alderman, E., and Kennedy, C. 1995. *The Right to Privacy*. New York: Knopf.

Alvey, W., and Kilss, B., eds. 1985. *Record Linkage Techniques—1985*. Washington, D.C.: Department of the Treasury, Internal Revenue Service.

Belin, T., and Rubin, D. 1993. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707.

Bellhouse, D. 1988. "A Brief History of Sampling Methods." In *Handbook of Statistics*. Elsevier Science. Representative sampling seems to have first been proposed in an 1895 paper by Kiaer. This was a major advance, as routine as such surveys seem today. Kish, in writing a paper on the history of surveys, captures the real battle (which continues) in the title of his presentation, "The Hundred Years War of Survey Sampling," which he delivered in Rome on May 31, 1995.

Blair, J. 1995. "Ancillary Uses of Government Administrative Data on Individuals: Public Perceptions and Attitudes." Unpublished Working Paper. Washington, D.C.: Committee on National Statistics, National Academy of Sciences.

Boruch, B., and Terhanian, G. 1996. "Trends in Statistical and Analytic Methodology: Implications for National Surveys." Paper presented at the Conference on Future NCES Data Collection: Some Possible Directions, Washington, D.C., November 1995, and included in this volume.

Boruch, R., and Cecil, J. 1979. *Assuring the Confidentiality of Social Research Data*. Philadelphia: University of Pennsylvania Press.

Brannigan, and Beier. 1995. *Medical Data Protection and Privacy in the United States: Theory and Reality*. Paper submitted with support from American Association of the Advancement of Science, Science and Human Rights Program.

Cecil, J. 1993. "Confidentiality Legislation and the United States Federal Statistical System," *Journal of Official Statistics* 9 (2): 519–536.

Czaijka, J., Moreno, L., and Schirm, A. 1995. *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population*. Report prepared for the U.S. Department of Treasury, Internal Revenue Service by Mathematica Policy Research, Inc.

Dalenius, T. 1988. *Controlling Invasion of Privacy in Surveys.* Continuing Education Series, Statistics Sweden.

Dillman, D. 1994. *Why Innovation is Difficult in Government Surveys.* Presented at the 1994 Census Bureau Annual Research Conference.

Duncan, G., Jabine, T., and de Wolf, V. 1993. *Private Lives and Public Policies.* Washington, D.C.: National Academy Press. See also the review by Scheuren, F. March 1995. *Journal of the American Statistical Association.*

Fay, F. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association.*

Gastwirth, J., and Johnson, W. 1994. "Screening With Cost-Effective Quality Control: Potential Applications to HIV and Drug Testing," *Journal of the American Statistical Association* 89: 972–981.

Ghosh, D., Smith, W., Chang, M., and Saba, M. 1995. *Optimizing the Periodicity of the Schools and Staffing Survey: An Interim Assessment Based on 1987–88 and 1990–91 Data.* Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

Granquist, L., and Kovar, J. 1995. "Editing of Survey Data: How Much is Enough?" Paper delivered at the International Conference on Survey Measurement and Process Quality, Bristol, UK, April 1995. Incidentally, one way to partially avoid the problem of overediting is provided in Lawrence, D., and McDavitt, C. 1994. "Significance Editing in the Australian Survey of Average Weekly Earning," *Journal of Official Statistics* 10: 437–447.

Groves, R. 1995. "Challenges of Methodological Innovation in Government Statistical Agencies." *The Future of Statistics, An International Perspective.* Ed. Z. Kenessey. International Statistical Institute.

Hall, G. et al. 1985. *Alternatives for a National Data System on Elementary and Secondary Education.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Harris et al. 1993. "Harris-Equifax Privacy Survey" (for 1993 and earlier). New York: Louis Harris and Associates.

Holt, A., and Scanlon, B. 1994. *QED Estimates of the 1990–91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Jabine, T. 1993. "Statistical Disclosure Limitation Practices of United States Statistical Agencies," *Journal of Official Statistics* 9 (2): 427–454.

Jabine, T., and Scheuren, F. 1985. "Goals for Statistical Uses of Administrative Records: The Next Ten Years." *Journal of Business and Economic Statistics*.

Kaufman, S., Li, B., and Scheuren, F. 1995. "Improved GLS Estimation in NCES Surveys." Paper delivered at the meetings of the American Statistical Association, Orlando. Incidentally, this paper forms a partial pilot for the poststratification ideas discussed here.

Kaufman, S. 1995. *Properties of the Schools and Staffing Survey's Bootstrap Estimator*. Paper presented at the meetings of the American Statistical Association, Orlando.

Keller, W. 1995. "Changes in Statistical Technology." *Journal of Official Statistics* 11 (1): 115–127.

Keller-McNulty, S., and Unger, E. 1993. "Database Systems: Inferential Security." *Journal of Official Statistics* 9 (2): 475–500.

Ligon, G. 1996. *New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education*. Paper presented at the Conference on Future NCES Data Collection: Some Possible Directions, Washington, D.C., November 1995, and included in this volume.

Malec, D., and Sedransk, J. 1995. "Small Area Inference for Binary Variables in the National Health Interview Survey." Washington, D.C.: U.S. Department of Education, National Center for Health Statistics.

McMillen, M., Kasprzyk, D., and Planchon, P. 1993. "Sampling Frames at the United States National Center for Education Statistics." *Proceedings of the International Conference on Establishment Surveys*, 237–243.

Mulrow, J., and Scheuren, F. 1995. *Measuring to Improve Quality and Productivity in a Processing Environment*. Paper presented at the Orlando meetings of the American Statistical Association. This is an application of Japanese, or Deming-like ideas, to surveys—focusing on cultural and organizational change.

Nanopoulos, P. 1995. "Expected Changes in Record Keeping," In *The Future of Statistics, An International Perspective*. Ed. Z. Kenessey. International Statistical Institute.

National Academy of Science. Committee on Applied and Theoretical Statistics. 1992. *Combining Information: Statistical Issues and Opportunities for Research*. Republished in *Contemporary Statistics* #1, American Statistical Association, 1992. This is a good general reference to the broad area of combining information. Many of the hybrid ideas covered in the section on Survey Investment Opportunities in this paper fall under this general rubric. Virtually every federal agency involved in assessing the impact of alternative policy proposals uses models depending on micro-data: information relevant to highly aggregate (e.g., national) conclusions, but measured at a much lower level of aggregation.

National Academy of Science. 1993. Student Aid Program.

Nevada Department of Education. 1994. *S.M.A.R.T. Plan, Statewide Management of Automated Record Transfer: A Plan to Automate and Transfer Student Records Statewide.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Newcombe, H., Fair, M., and Lalonde, P. 1992. "The Use of Names for Linking Personal Records," *Journal of the American Statistical Association* 87: 1193–1207.

Newcombe, Howard B. 1988. *Handbook of Record Linkage*. New York: Oxford University Press.

Newmann, F., and Wehlage, G. 1995. *Successful School Restructuring*. A Report to the Public and Educators by the Center on Organization and Restructuring of Schools, University of Wisconsin, School of Education.

Nicholls, W. 1988. "Computer-Assisted Telephone Interviewing: A General Introduction," In *Telephone Survey Methodology*. Eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg. New York: Wiley, pp. 377–402. See also Batcher, M. and Scheuren, F. 1996. CATI site management in a survey of service quality, *Bristol Survey Quality Measurement Conference*. Wiley: New York. In the literature on computer-assisted surveys, usually the observation is made that data capture costs do not drop over what were, heretofore, the conventional methods. Because of the higher quality obtained, however, it could be argued that total costs have fallen.

Noam, E. October 1995. "Electronics and the Dim Future of the University," *Science* (270): 247–249. The university has been based on a "model—centrally stored information, scholars coming to the information, and a wide range of information subjects housed under one institutional roof—logical when information was scarce, reproduction of documents expensive and restricted, and specialization low. It became also the model for the most formidable of knowledge institutions of antiquity, the Great Library of Alexandria." These conditions are no longer true and, as a consequence, major changes seem in store.

Office of Science and Technology Policy. 1994. *High Performance Computing and Communications: Toward a National Information Infrastructure*. A Report by the Committee on Physical, Mathematical and Engineering Sciences.

Olmstead v. United States. 1928. 277 U.S. 438–478. (Justice Brandeis dissenting).

Peng, S., Gruber, K., Smith, W., and Jabine, T. 1993. "Monitoring Data Quality in Education Surveys." *Proceedings of the International Conference on Establishment Surveys*, pp. 244–252. Some of these are also discussed in the NCES *Working Paper Series*. Washington, D.C.

Presser, S., and Singer, E. 1995. "Public Beliefs about Census Confidentiality." Preliminary Report based on part of the 1995 Joint Program in Survey Methodology Practicum Survey.

Purcell, N., and Kish, L. 1979. "Estimation for Small Domains," *Biometrics* 35: 365–384.

Rennie, J. September, 1995. "The Uncertainties of Technological Innovation." *Scientific American, Key Technologies for the 21st Century, 150th Anniversary Issue.* According to Rennie, "Few of the promised technologies failed for lack of interest." One recurring reason that predictions were off "is that even the most knowledgeable forecasters are sometimes much too optimistic about the short-run prospects for success." He then adds, "The more fundamental problem with most technology predictions . . . is that they are simplistic and, hence, unrealistic. Sadly, some inventions are immensely appealing in concept but just not very good in practice."

Rubin, D. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association.*

Rubin, D. Discussant remarks in the *1990 Proceedings of the Annual Research Conference, (pp. 676–679).* Washington, D.C.: Bureau of the Census.

Salvucci, S., Zhang, F., Monaco, D., Gruber, K., and Scheuren, F. 1995. "Multivariate Modelling of Unit Nonresponse for the 1990–1991 Schools and Staffing Surveys." Paper delivered at the American Statistical Association Meetings, Orlando.

Särndal, C.E., Swensson, B., and Wretman, C. 1991. *Model Assisted Survey Sampling.* New York: Springer-Verlag. For a criticism of such methods for a statistical agency, see Smith, T. 1994. "Sample Surveys: An Age of Reconciliation, 1975–1990." *International Statistical Review* 62: 5–34.

Schafer, J. 1991. "Algorithms for Multiple Imputation and Posterior Simulation from Incomplete Multivariate Data with Ignorable Nonresponse." Unpublished Doctoral Dissertation, Department of Statistics, Harvard University.

Schafer, J., Khare, M., and Ezzati-Rice, T. 1993. "Multiple Imputation of Missing Data in NHANES-3." *Proceedings of the Annual Research Conference* (pp. 459–487). Washington, D.C.: Bureau of the Census.

Schaible, W. ed. 1996. *Indirect Estimators in U.S. Federal Programs.* New York: Springer-Verlag.

Scheuren, F. 1985. "Methodological Issues in Linkage of Multiple Databases." *Record Linkage Techniques-1985.* Washington, D.C.: Department of the Treasury, Internal Revenue Service.

Scheuren, F. May 1995. "An Administrative Record Census in the U.S.?" *Chance*. The costs of primary data collection for statistical purposes are simply not competitive with an administrative source, if such a source can be made to serve. Why capture the same, or almost the same, data more than once? This point has been made over the last 15 years with regard to census taking. The 2010 Decennial Census will almost certainly be based in large part on administrative records, as is already common for many European countries. Education records could be among those used; much will depend on how the 2000 Census research program is carried out.

Scheuren, F. March 1995. "Review of Private Lives and Public Policy." *Journal of the American Statistical Association.*

Scheuren, F. 1995. *Linking Health Records: Human Rights Concerns.* Paper submitted with support from American Association of the Advancement of Science, Science and Human Rights Program.

Scheuren, F., and Li, B. 1995. *Intersurvey Consistency in NCES Private School Surveys.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Science 1995. "Getting Wired." *Random Samples* 270: 1929.

Singer, E., Shapiro, R., and Jacobs, L. 1995. "Privacy of Health Care Data: What Does the Public Know? How Much Do They Care?" Paper submitted with support from the American Association for the Advancement of Science, Science and Human Rights Program.

Singer, E. 1993. "Informed Consent and Survey Response: A Summary of the Empirical Literature." *Journal of Official Statistics* 9 (2): 361–375.

Spruill, N., and Gastwirth, J. 1982. "On the Estimation of the Correlation Coefficient from Grouped Data." *Journal of the American Statistical Association* 77: 614–620.

Toffler, A., and Toffler, H. 1994. *Creating a New Civilization.* Atlanta: Turner Publishing. The authors provide a set of questions that allow one to determine whether an approach is "third wave." Among other things, a third wave approach should 1) not resemble a factory—i.e., not be built on such principles as standardization, concentration, and bureaucratization; 2) not massify society—i.e., focus on the individual; 3) not have too many eggs in one basket.—i.e., push as many decisions as possible down from the top and out to the periphery; and 4) not be vertical but be virtual. Contrast this with the centralized vertically integrated survey system proposed for NCES in Hall, G. et al. 1985.

U.S. Department of Education. 1993. *SpeedeExpress, An Electronic System for Exchanging Student Records.* Washington, D.C.: Council of Chief State School Officers, National Center for Education Statistics.

U.S. Department of Education. 1995. *Advanced Telecommunications in U.S. Public Schools, K–12*. Washington, D.C.: National Center for Education Statistics.

U.S. Department of Education. 1994b. *Education Data Confidentiality: Two Studies, Issues in Education Data Confidentiality and Access and Compilation of Statutes, Laws, and Regulations Related to the Confidentiality of Education Data*. Washington, D.C.: National Center for Education Statistics, National Forum on Education Statistics.

U.S. Department of Education. 1994a. *The Condition of Education 1994*. Washington, D.C.: National Center for Education Statistics. This publication may be a direct response to the need to monitor progress toward the goals mentioned in the report *Education Counts: An Indicator System to Monitor the Nation's Educational Health*. 1991. Moreover, the *Condition of Education* report series is, of course, only one of many general and specialized NCES publications that illuminate this important national priority.

U.S. Department of Education. 1994c. *Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study*. Washington, D.C.: National Center for Education Statistics. For a somewhat different perspective, see Rotberg, I. December 1995. "Myths about Test Score Comparisons." *Science* 1: 1446–1448.

U.S. Department of Education. 1991. *Education Counts: An Indicator System to Monitor the Nation's Educational Health*. Report of the Special Study Panel on Education Indicators to the Acting Commissioner of Education Statistics. Washington, D.C. Among the goals to be monitored for the year 2000 are 1) All children in America will start school ready to learn; 2) The high school graduation rate will increase to at least 90 percent; 3) American students will leave grades 4, 8, and 12 having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography (and leave school) prepared for responsible citizenship, further learning, and productive employment in the modern world; 4) U.S. students will be first in the world in science and mathematics achievement; 5) Every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship; and 6) Every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning.

U.S. Federal Committee on Statistical Methodology. 1995. "Electronic Dissemination of Statistical Data." *Statistical Policy Working Paper 24*. Washington, D.C.: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

U.S. Federal Committee on Statistical Methodology. 1990. "Data Editing in Federal Statistical Agencies." *Statistical Working Paper No. 18*. Washington, D.C.: Office of Management and Budget.

U.S. Federal Committee on Statistical Methodology. 1980. *Report on Exact and Statistical Matching Techniques*. Washington, D.C.: U.S. Department of Commerce.

U.S. Information Infrastructure Task Force. 1995. *Privacy and the National Information Infrastructure: Principles for Providing and Using Personal Information.* A Report of the Privacy Working Group. Washington, D.C.

Westin, A. 1967. *Privacy and Freedom.* New York: Atheneum.

Whitridge, P., Bureau, M., and Kovar, J. 1990. *Mass Imputation at Statistics Canada.* U.S. Bureau of the Census, Sixth Annual Research Conference.

Winkler, W. 1995. "Matching and Record Linkage," in *Business Survey Methods.* Eds. Cox, Binder, Chinnappa, Christianson, Colledge, and Kott. New York: Wiley.

Winkler, W., and Scheuren, F. November 1995. "Linking Data to Create Information." Paper delivered at the 12th Annual Statistics Canada Methodology Conference, Ottawa.

Wright, D., and Ahmed, S. 1990. "Implementing NCES's New Confidentiality Protections." *1990 Proceedings of the American Statistical Association,* Section on Survey Research Methods. Alexandria, Va.

Wong, W., and Ho, C. 1991. "Bootstrapping Post-Stratification and Regression Estimates from a Highly Skewed Distribution." *1991 Proceedings of the American Statistical Association,* Section on Survey Research Methods. Alexandria, VA.

APPENDIX

National Postsecondary Student Aid Study by Dennis Carroll

In 1996, the fourth National Postsecondary Student Aid Study (NPSAS) will collect information from all types of students in all types of postsecondary institutions. This study, which is about two-thirds the size of previous administrations of NPSAS (1987, 1990, and 1993), builds upon the collection strategies of the earlier studies and incorporates administrative records from the major student financial aid programs. NPSAS is a comprehensive study, spanning aided and unaided students, independents and dependents, employed and non-labor force participants, undergraduates and students seeking advanced degrees, as well as full-time and part-time students. There are three major users of NPSAS data: NCES, USED, and financial aid policy analysts. NCES uses NPSAS to profile groups of students (e.g., undergraduates, part-timers, minorities, borrowers), and NPSAS serves as the base year for longitudinal studies (i.e., BPS and B&B). USED (including PES and OPE) used NPSAS to determine the rates of receipt of federal student financial aid for various subgroups of students. Policy analysts describe aid issues and build models.

NPSAS is an extremely complicated set of six integrated data collections: enrollment lists, Central Processing System (CPS) records, Computer Assisted Data Entry (CADE) for institution records, Computer Assisted Telephone Interviews (CATI) for students, CATIs for parents, and Pell grant/loan award files. From a sample of all (IPEDS) institutions, lists of all students (undergraduate and graduate/first-professional) enrolled at any time during the academic year (July 1–June 30) are collected (with business majors flagged, if possible). For the longitudinal component, lists of first-time students (BPS) and filed for graduation students (B&B) are collected and unduplicated. The initial NPSAS sample of students is selected from these lists.

The USED CPS contains application and preliminary award information for federal student aid, in particular, the Pell grant program. The initial NPSAS sample is matched with the CPS information to obtain data on family finances and preliminary awards of federal aid. (CPS does not contain all federal aid. Much of the federal loan program data are currently fragmented in several files, which may be consolidated in the next several years.) Matched CPS data are preloaded into the CADE.

To extract data from institutional records housed in student financial aid offices, admission offices, and graduate dean's offices, a CADE is used. Many institutions (over 60 percent) complete the CADE data collection on their own, but some institutions do not have staff or computers. Contractor staff travel to these institutions and complete the CADE for them. The CADE extends CPS data to gather the core NPSAS information on student aid—including all federal aid, state aid, institutional aid, and assistantships. In addition, information on program (e.g., intensity, major, admission, and demographics) is collected. The CADE data are preloaded into CATI systems.

The CADE data allow subsampling of students who did not receive student financial aid. (About 40–45 percent of students receive aid.) In NPSAS:96, the initial sample of 59,000 has been reduced to a more efficient sample of 37,000 by undersampling unaided students.

The two CATI systems for students and some parents require location and collection systems. Only a subgroup of parents are interviewed consisting of mostly dependent and unaided students' parents. If in tracing the student the parent is contacted, then the parent CATI is conducted; otherwise, the parent CATI follows the student CATI. The parent CATI gathers data on family finances that parallel the CPS information.

The student CATI expands the NPSAS data to cover six areas: other aid, non-school costs, labor force activities, family structures and finances, future plans and goals, and community activities. Other aid covers small programs that do not flow through student financial aid offices and aid from other institutions attended during the academic year. Non-school costs include living expenses, transportation, and child care. Labor force activities include employment (sometimes in college work-study program jobs), program related employment (internships), and lack of employment. Family structures and finances include marriage, children, and other dependents; the earnings/assets of the household; and the expenses of the household. Future plans and goals include occupational, community, and personal aspirations. Community activities include citizenship and service.

Finally, the Pell grant and federal loan award files (in their final audited form) are merged.

NPSAS yields three recurring policy reports, two data analysis systems, and a restricted set of data files for secondary analyses. The recurring policy reports are *Profile of U.S. Undergraduates, Financing Undergraduate Education*, and *Financing Graduate/First-Professional Education*. Separate data analysis systems are built for undergraduates (about 700 variables) and graduate/first-professionals (about 600 variables). Finally, the data files (with the associated methodology report) are made available to licensed users for secondary analyses (including Postsecondary Education Descriptive Analysis Reports [PEDAR]).

New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education

---

Glynn D. Ligon

OVERVIEW

The National Center for Education Statistics (NCES) is seeking a future vision for data collection, storage, retrieval, and reporting. This vision will guide improvements in data collection and reporting processes to increase the availability and usefulness of data while decreasing the burden on local and state agencies. This paper describes the developments in technology that will affect the collection and reporting of education data. A major implication is that information solutions present challenges that are as much human resource issues as technology issues. The lack of acceleration in our use of technology is attributable in large part to the shortage of individuals trained and capable of making the technology work, within an environment that encourages the use of technology. For NCES, staffing roles, responsibilities, and skills must change along with the introduction of technology solutions.

Summary of Implications for NCES

NCES should position itself to ride the wave of automation in the nation. The trends described herein are as follows:

1) Faster computers will allow NCES to expand the amount of data collected, analyzed, and reported while potentially reducing the time and burden imposed on reporting agencies and NCES staff.
2) Increased storage capacity on computers will allow NCES to collect and maintain as much data as is reasonable to collect based upon the information needs of audiences.
3) The universality of networks will allow NCES to collect data electronically, communicate to clients electronically, and make available its analyses and reports electronically.
4) EDI standards and software will make electronic data exchanges over these networks efficient, effective, and affordable.
5) Relational data base concepts will be applied to a distributed information system that will allow access to data across agencies' files.

6) Productivity software will automate information management tasks to the extent that staff will insist upon computer applications over any remaining manual processes.

7) NCES can achieve the benefits of an individual student-level database without the problems of creating one within NCES. Emerging networks and data standards can create a national distributed information system. NCES would be able to query each state database to conduct analyses without having to maintain individual records centrally.

The NCES Data Warehouse

Technology supports NCES's plans to develop a data warehouse. A data warehouse is simply a location where someone can access information electronically. The NCES data warehouse should be a library containing both books with statistics and analyses already accomplished and raw data available for analysis.

Criteria for Judging the Future System

"Alternatives for a National Data System on Elementary and Secondary Education," 1985, proposed a set of criteria to be used for judging a national education information system. These criteria are applied to the vision described here. In a reverse of position from 1985, confidentiality will move from the bottom to the top of the list of concerns requiring careful attention by NCES.

Conclusion

Ensuring that NCES's data collection, storage, analysis, and reporting processes take full advantage of technology will be a process, not an event. This transition will require considerable training and support for both NCES staff and the staff of its data providers. When evaluated against the criteria described in 1985, the vision of the future as described here would be a significant improvement over past and current systems.

INTRODUCTION

The National Center for Education Statistics is seeking to establish a vision for data collection, storage, retrieval, and reporting for the future. This vision will guide the planning and implementation of improvements in the data collection and reporting processes. This effort is significant for many other agencies beyond NCES. Nationwide, decision makers, parents, educators, students, businesses, and others are affected by the availability and quality of education data. The expectation is that technology advances will provide opportunities for solutions that will increase the availability and usefulness of data while achieving decreases in the burden imposed upon local and state agencies to collect and report the data.

Technology is already available to support the processes described. In fact, the NCES staff have already used some of the newer methodologies on a limited scale. A challenge will be to escape the inertia of traditional systems, to create a new inertia of change, one that shortens the time required to go to scale with technology-enhanced solutions."

This paper describes the developments in technology that have affected or will affect the collection and reporting of education data. The underlying premise is that we must have a vision for a new national education information system. Our vision, based upon function, needs to drive our decisions and actions. At present, many decisions are reactions to new technology as it is developed. We are intrigued by technology and want to adapt our needs to it. An important perspective for us in the education information arena is that our needs should inspire a search for technology that provides solutions to those needs. The functional aspects of our data collection and reporting should change as our needs change. Technology is one direction in which to look for solutions to our changing needs.

A major implication from the discussion in this paper is that the information solutions to be explored and implemented present challenges that are as much human resource issues as technology issues. For education institutions, the reality has been that the capability of technology is ahead of the capability of individuals to apply that technology to our information systems. In other words, much of what educators are asking to do now can be accomplished with existing technology or straightforward adaptations of hardware and software. The lack of acceleration in our use of technology is attributable in large part to the shortage of human resources, individuals trained and capable of making the technology work, within an environment that encourages the use of technology. For NCES, staffing roles, responsibilities, and skills must change along with the introduction of technology solutions.

Importantly, the future system described in this paper must be responsive to issues raised in the other papers in this series. The technology used in our future information systems must be chosen because it is responsive to the demands detailed in the other papers. A major concern within education agencies today is the purchasing of hardware and software because they are available and appear to be useful. The alternative is to seek hardware and software solutions for problems that have been clearly identified. This is particularly evident in the instructional arena where the users may be inspired by technology, but not have the time and resources to integrate it adequately into their processes.

For NCES and the future of education information at the national level, a major hurdle with which to contend is the variety in both type and age of the technology that must be integrated across schools, districts, postsecondary institutions, states, and NCES to create a functional information resource for decision makers, parents, businesses, educators, staff, and others. Schools, districts, and state agencies have been acquiring technology (e.g., computers, printers, modems, and so on) since the early 1980s. Much of that hardware is still in use, irrespective of how out-of-date it has become. Some states purchased hardware when large sums of dollars became available to their legislatures. Since that time, dollars to upgrade have been more difficult to find. So, on the one hand, this paper makes the point that hardware is relatively inexpensive to purchase now. However, on the other hand, available funds for purchases may be scarce.

As examples, consider the situations in South Carolina and Georgia. South Carolina raised taxes for education once about a decade ago and purchased that era's state-of-the-art hardware and a student information management software system for schools. Now the software is being updated by the vendor, requiring a newer, more powerful operating system that will not run on the old hardware. About 7 years ago, the Georgia Legislature approved funds to build a student information system across all schools. Over those years, hardware purchases have been made to bring about 70 percent of the schools to an operational status. However, those schools where implementation occurred years ago have old hardware compared to the schools being brought on board this year.

Another perspective is found in Texas. Almost a decade ago, plans to build a statewide information database were begun. At the time, accommodations were made in the design for schools that were still punching 80-column cards. Currently, the 80-column format is still being used for the computer files submitted. Other states are designing information systems now that are incorporating relational database designs to be much more efficient. The dollars and human resources required for Texas to reengineer its existing system are huge.

Technology may be capable, but are the users in education agencies ready? Is the technology present in the education arena? How out-of-sync will agencies become as the financially advantaged acquire capabilities that others do not have—or as agencies replace and upgrade at varying rates? Do we have to plan for the lowest common denominator?

We need to build a vision of functions, not of hardware. We need to envision systems whether or not the infrastructure exists to support them, then we need to build toward that vision, ensuring that each step taken is consistent with the long-range goal. Space travel has taught us that all the pieces of technology do not have to be in place before a project can begin. New techniques and products can be developed along the way.

Data Collection

Data quality must be achieved and maintained throughout all areas of information systems, but it begins with adequate standards during collection. There is a balancing between timeliness and quality that threatens to undermine the ultimate purpose of data collection, which is to inform decision making. Data quality must become a priority for the future. Information systems must be designed to provide timely education data that can be used with confidence as the basis for decisions.

The mechanics of data collection are changing already—from paper-and-pencil forms to optical scanners, to computer screen entry, to disk exchanges, to electronic file transfers, to direct reading of distributed files, to simultaneous updating of remote files as transactions occur. All of these are existent to some degree across educational agencies. Our vision must motivate agencies to continue moving up the hierarchy of automation. As data are collected at the local and state levels in automated fashion, they are more readily available for exchanging up the system to NCES. The vision for data collection must include the idea that redundant, independent data collections will be coordinated. Changes in retrieval and access processes as described below allow for collection of data from files within databases rather than requiring that someone reformat the data to fit a forms-based report.

Data Storage

Data storage media are increasing in capacity and decreasing in cost. The changing formula of cost-per-piece of information stored indicates that we can and will allow ourselves to be less disciplined about what we store and how long we maintain it. The implication is that more unrefined, raw data will be maintained and be available for analyses. With faster processors, more individuals will have the ability to process huge files of raw data. Already we have seen the discipline of sampling theory decline in importance in research. More studies are conducted on population statistics rather than sample statistics, because the cumbersome calculations required for sophisticated statistical procedures are handled easily by computers. Advances have changed how we store images, translate voice to text, scan text and translate it to word processing files, and create documents and data files without ever producing a paper document. This paper discusses the implications for coping with a data system that grows to include so many elements in so many formats. The emerging methodology for data warehouses will provide some answers.

Data Retrieving

Retrieving, which will also be thought of as access, is the function that supports the utility of data and makes it more valuable. In the automated world, the separate concepts of retrieving and disseminating begin to blur. As audiences gain access to data, the act of someone disseminating the data is no longer necessary. Retrieving and disseminating can be viewed as all being part of a single process that makes data available to users in a wide range of states of development from raw to fully analyzed. This paper discusses how future information systems will employ a range of access techniques to accomplish retrieval/dissemination. Access will be closely linked to issues of security, confidentiality, and integrity of both the data themselves and the analyses and conclusions drawn from them. This will be a controversial issue. Determining who can access which data elements within a database will be difficult. Controlling access to ensure that only those authorized to access certain data are allowed to will be an even bigger challenge.

"Regulated access" allows the owner of data to place them in a location for access, without requiring that owner to package and send them to every requester. Today, someone within an organization typically prepares a response to an information request and sends it. With regulated access, the owner of the data will monitor who is accessing and using them rather than providing the data directly. Requesters/readers have responsibility for establishing their credentials for access and usage.

A CHANGE IN PARADIGMS

Although overused today, the phrase "changing our paradigms" applies precisely to the automation of data collections and the use of the resultant data. An important concept in this change will be that the nature of data collection will evolve. We will not want to merely automate manual or paper systems. When conversions are made to technology-based systems, the process underlying the collections should change to take full advantage of how the technology operates.

Some other ways our thinking must change are as follows:

1) Survey forms *will be replaced by* data files that do not look at all like the paper surveys.

2) Dissemination of reports *will be replaced by* interested audiences accessing information in electronic form and printing the parts they want.

3) Statistics calculated and published by a single agency *will be replaced by* competing statistics calculated from the same database by both private and public entities.

4) Keeping all the data you need on your own computer *will be replaced by* networked databases from which your computer can access huge data sources.

5) A computer programmer responding to a request for a report *will be replaced by* having the person who needs the report run it.

6) New mandates requiring new data collections *will be replaced by* new mandates resulting in an analysis of data from an existing, shared database.

7) Data burden being defined as the amount of time required to document activities and complete reports *will be replaced by* its being defined as the overwhelming amount of data available for consideration.

8) Statistics and reports being published months after collection *will be replaced by* immediate access to data as soon as they are uploaded to a central file.

Within the context of its charge to provide useful and timely statistics about education, NCES is finding that many other agencies and organizations collect and report data as well. Professional organizations survey members and the general public often these days. Commercial polling services conduct numerous, seemingly continuous, surveys of public opinion. With the expansion of computer storage capacity and the move toward providing public access to data and report files, there arises the issue of how much of these related data collections should be acquired and made available by NCES.

Several issues are clear. First, does NCES endorse or make an implicit statement about the quality of other organizations' data by redistributing them? What obligation does NCES inherit when it redistributes these data? Secondly, is this redistribution necessary? As will be described in this paper, the technology allows for NCES to point audiences to other information sources using electronic connections without having to copy the data they are seeking onto an NCES computer. The cautious approach would be to leave data collected by other organizations and agencies on their own information systems and resolve the technology issues of how to connect potential audiences to them as appropriate.

WHAT ARE THE DEVELOPMENTS IN TECHNOLOGY THAT AFFECT EDUCATION DATA SYSTEMS?

Advances in technology are very technical and complex within the covers of our computers and other hardware. However, to the users of information systems, the relevant aspect of these advances is function. Function can be described as the operational actions that a user notices. What does the application do for you? How well does it do it? How fast does it perform that function? What manual activities are replaced? When microprocessing chips are miniaturized through amazing advances in manufacturing, the end user notices that computers grow smaller and faster at the same time. When modems advance in their transmission speeds, the end user notices that activities that used to take too long to be practical over a phone line can now be accomplished reasonably. So, the technical advances that result in faster chips and modems are discussed here more in terms of the impact they have on users. The impact on users translates directly to implications for the next generation of NCES data collection and reporting systems.

Developments and their implications for NCES are discussed within these areas:

Hardware:	The physical items that make up the computer and its visible components
Network:	A group of two or more computer systems linked together; the telecommunications systems that link computers
Software and Applications:	The instructions that tell the computer what to do

Hardware

Storage Capacity

Compared to the 1980s, today's data storage devices present fewer limitations on the quantity of data we can have readily available to us. A storage device is the object onto or into which data are placed. These include hard disk drives (internal or external magnetic disks); removable floppy diskettes, cassettes, or cartridges using magnetic disks or tape; and optical disks (compact disks or CDs). For comparison purposes, commonly found hard drives of under 50 megabytes in the 1980s would not even hold some of today's data files that can exceed 100 megabytes for elaborate publications with graphical images. As this paper is being written, families are buying 1 gigabyte hard drives for their homes. The floppy disks of a decade ago have been replaced by removable disks and cartridges that hold several gigabytes of data.

Storage capacity is not limited by the advertised level on floppy disks, tapes, and cartridges. The demand for affordable, large storage has inspired software developers to design data compression routines that remove all the unnecessary bits of information out of a file. These compression routines can achieve impressive results, such as reducing the space required to store a file by 10 to 90 percent.

The impact of these advances in storage efficiency is that the limits are being removed on the number of files and the amount of data that can be maintained within an individual's computer. When NCES began keeping the many statistics it collects from the states on computer files, the size of those files was a major consideration, and the cost to add more storage to hold more files and data had real budget impact. Today, several hundred dollars can solve a large data storage need. The direction of technological advances continues to be toward greater and greater storage capacities, in less and less space, for fewer and fewer dollars.

For NCES, this means that constraints that used to be placed upon expansion of data files and conversion of paper records to an automated format have faded. NCES is capable of holding within local computers virtually all the data that are practical to collect and enter. Future decisions determining the data to be collected can be made upon need and usage factors rather than available storage capacity.

In the past, researchers were required to understand and use sampling theory to create reliable data sets for analysis. With limitations upon the ability to access data on mainframe systems or to store large data sets on personal computers, a premium was placed upon collecting manageable sample data sets. Considerable professional literature has been produced to guide researchers in this process. Probability statistics have been common in the literature to provide readers with an understanding of how much confidence they should place in the findings of studies. Educational research is now using population statistics from large databases that include measures of every individual of interest. The constraint is more on the collection methodology (how practical it is to measure every individual) than on the data storage and analysis capacities. This trend will be evident in the future operations of NCES. As a data warehouse is built and stocked, more and more data will find its way into it. Fewer and fewer restrictions will be imposed based upon lack of storage space.

Another aspect of data storage that has changed involves the benefits from expanded electronic networks. With a local area network installed, NCES can store data on multiple computers throughout the agency and create an environment that functions as a virtual single source for data. This concept also works on a much broader scale outside NCES. Any agency that shares a common set of standards for exchanging data files can be a part of a distributed information system. Such a system would allow sharing of data while maintaining internal integrity and local control. This would be in contrast to a true distributed database within which all agencies must comply with exactly the same data definitions and formats. Those implications and benefits are discussed in more detail throughout this paper. The bottom line is that in a networked environment, the users have virtually on their own desktops the data from all computers linked by the same network.

Telecommunications Speed

In the 1950s, "faster than a speeding bullet" (miles per hour) impressed us because it was too fast for us to actually see. In the 1990s, data traveling virtually at the speed of light carries our communications over fiber optics. The result is that we no longer describe the efficiency of a computer or the transmission of data as "how fast something is moving." Our data transmissions have reached a plateau in how fast they move. Speed is now defined in terms of how much information can be sent from one place to another within a certain amount of time (bits per second). The bullet Superman outraced traveled intact, arriving at its destination in the same physical shape as it left. Data files are stretched out and arrive literally in bits and pieces. In telecommunications, the goal is to send and receive as many data in as little time as possible. In other words, the performance goal is to stretch out the data file as little as possible so the first bit that arrives is followed as soon as possible by the last one. This goal has been pursued with great success. Most casual personal computer users have noticed that their modems (the devices that translate information into and out of the characteristics required for transmission) evolved very quickly from 1200 baud (roughly 1,200 bits per second) to 2400, 9600, 14.4 (notice the change in notation to units of 1,000 with 14,400 being expressed as 14.4), to a common modem on store shelves transmitting at 28.8. A 28.8 baud modem sends about 24 times as much data as the old 1200 baud modem did in the same amount of time. This miracle is achieved in great part through eliminating any unnecessary bits of information in a data file and compressing everything into as few bits as possible to carry the same meaning when decompressed at the other end.

What implications does speed have for the future of NCES data collection and reporting? Faster telecommunications will allow for larger data sets to be exchanged efficiently. Again, this removes a barrier to designing future systems. Future information systems will not have to be constrained as much by the time and expense factors in data exchange. NCES can collect more data in large data sets without imposing a greater burden on states and others in terms of transmission time and costs. Today, NCES's trading partners are already finding it to be more practical to extract and transmit data electronically compared to copying data onto a floppy disk and physically sending it. Across the state education agencies, few have not implemented some data submissions on disk, and some have implemented submissions over networks.

Processing Speed

Another speed issue is how long a computer takes to perform the millions of transactions it is asked to do for a specific application. Processing speed is one of the more difficult concepts to discuss. There are many factors that determine actual processing time for a computer task, e.g., access time for storage devices, input/output time for other components of the computer system, and the amount of time required for the monitor to recreate images as they change. Even the casual personal computer user knows that the speed at which personal computers' central processing chips perform tasks has increased dramatically. Miniaturization in the manufacturing of chips continues to progress. Simply put, tasks that took hours in the 1980s were reduced to minutes in the 1990s, and are now being completed in seconds.

The implication for NCES is similar to that for all education researchers. We can now calculate complex analyses on large data sets within a more reasonable amount of time. As discussed earlier, the need for sampling strategies and sampling statistics is reduced. A researcher can use an entire data set on a population of individuals. For future planning, NCES does not have to be so concerned with having large data sets to analyze and the burden that places on staff and the time that requires to publish statistics. The option presented to NCES will be to produce more and more analyses and reports within the same amount of time, or to publish the same analyses more quickly—or both.

Access Speed

When speed is discussed, there is another component beyond the central processor and the modem that has an impact on how quickly an individual can accomplish work on a computer. A major factor is access speed for all the storage devices. The access speed determines how long it takes the computer to move data from the storage device into its active memory (random access memory or RAM). Data must be in RAM to be processed. Larger computer programs require that data be moved into and out of RAM periodically. CD-ROM players moved from single-speed access to 4x, or four times, the speed for the original cost within about 2 years.

Improvements in access speed contribute to the overall performance of computer systems. Again, the limitations on future information systems of NCES are shrinking. The task of maintaining and using a very large information system is taking less time.

Random Access Memory (RAM)

RAM is the random access memory a computer uses to keep data readily accessible for processing. A useful analogy is the human brain. The brain stores tremendous amounts of memories. We could never keep all those memories active in our conscious at one time, so only that information that is needed for thinking at any one time is called upon. The computer calls up those data it needs for the current task it is performing into RAM. The greater the capacity of a computer's RAM, the more information that can be kept handy for processing at one time. Commonly installed RAM has grown from 1 to 2 megabytes 3 years ago to 4 to 16 today. Newer operating systems (the essential directions that tell the computer how to run software programs) require greater RAM. This trend appears to be a given to continue or even to accelerate.

Another counterbalancing trend is the increased usage of RAM and storage capacity by newer operating systems. The implication of this is that as operating systems (e.g., Windows 95) improve, they will require more RAM to operate and more disk space to be loaded. The future of prices for RAM is uncertain, so it is not possible here to predict whether the increased RAM required in the future will cost more than the amount required in today's machines.

There is also a benefit for large information systems. Computers with adequate RAM will perform large, complex tasks quicker. This is one area where added productivity comes at a cost. The installed computers in many offices are old enough to have inadequate RAM to run the newest operating systems and applications.

Printers and Graphics

A brief note is appropriate here about the visual appeal and communicability of the output from the newest publishing/printing systems. A desktop computer can now produce the impressive color graphics that once were the sole venue of professional layout artists and printers. For NCES, the benefit is that staff can make publications more reader-friendly and more likely to be read.

Network

Up front, we should recognize that easier access to networks has been a priority feature of newer operating systems (e.g., OS/2 Warp, Windows 95). The user's challenge to learn how and to take advantage of networks becomes easier with each new generation of operating systems.

Local Area Networks

Computers within a single location can be connected to each other to share resources in a local area network (LAN). Physically, a LAN consists of a card inside each computer, wires between the computers, and network software to manage the communications between the computers. Printers and other devices may also be connected through the LAN. Some LANs are this simple. Others can use wireless communications, multiple access units and routers to direct the transmissions between locations, and servers. Servers are computers that store data and software, and manage the operations of the LAN.

LANs expanded in the late 1980s as users discovered the advantages of sharing printers, using electronic mail, working on the same documents, and reading data on another computer. A single user gained the power of several computers. With the recent installation of a LAN within NCES, this potential is available to staff.

Wide Area Networks

The Internet is a wide area network (WAN). WANs connect computers that are located in separate places. LANs may be connected by WANs. The distinction between a LAN and a WAN is the amount of separation between the computers. However, the technical requirements, legal parameters, and operational issues for a WAN are much more complex than for a LAN that is self-contained within a single location.

The Internet is a public network that connects anyone to anyone else who chooses to connect. Public institutions, including state education agencies and postsecondary institutions, are almost universally connected. Across these agencies, the level of usage varies. However, NCES currently has access to its major data trading partners through the Internet. School districts and schools are connecting quickly. However, some are far from being automated in their operations, and some of those choose not to be for the foreseeable future. Therefore, NCES can assume that the Internet is available for use by its primary information trading partners, but that those partners may be exchanging information with others who are not connected to the Internet.

Although not free as is commonly thought, the Internet is relatively inexpensive to connect to and use. The Internet is far from simple to access within some agencies. In 1995, NCES and the Office of Migrant Education sponsored a pilot across six sites to use the SPEEDE/ExPRESS standards as a basis for the exchange of education records for Migrant Education Program students. The expectation was also that the solution for migrant students would apply as well to all mobile students, who make up about 20 percent of students annually. Each volunteer site was to be connected to and using the Internet as a prerequisite for participation. The reality was that one site had only personal accounts used by a few staff members, another had no connection, another was connected, but required a multistep process for the Migrant Program staff to be trained and issued an address, and another used a gateway to a university Internet provider that required changes in the EDI software being used to connect. The other two sites were in Florida, which has an established statewide network. However, the Internet connection was set up through their state-level office rather from each district. The pilot demonstrated that the logistics of actually using the Internet for data exchanges can be much more involved that some may think.

Value Added Networks

Value added networks (VANs) are the private enterprise equivalent of the Internet. Although structured very differently, the functionality of VANs and the Internet are similar. Customers pay a VAN for usage of their network services. The value-added aspect is that the VAN provides services and features that the Internet expects the individual users to take care of themselves. The features include controlling access to users, guaranteeing connections, and providing some degree of security.

Very recently, VANs began making connections to the Internet available to their clients on a limited basis. Although too early to count on, the trend is for VANs to create more transparent connections with the Internet and to develop methods for maintaining the security and reliability that have been the key value-added features that have attracted users. VANs will be very cautious about risking their hard-won reputations for security by connecting to the public Internet. Stories are publicized frequently as another computer buff figures out how to break the code underlying current security and encryption techniques.

For NCES, one issue is the selection of the WAN to use. If indeed VAN-to-Internet connections become universal and functional, then NCES, as all other users, will be able to select the WAN or WANs that meet their needs the best. In the short term, the Internet's universality among public agencies and growing corps of proficient users argues strongly for its prominence in any planning.

Direct-Dial Connections

An alternative to these networks is a direct connection between two computers. A VAN or the Internet is not required to connect computers. The telephone companies provide connection using regular voice lines. One computer can dial another directly through their modems. This option provides for higher levels of security. Users can be required to have passwords for identification. Systems can also be set up to receive a call, then dial the caller back to ensure that your computer is really talking to the one identified as the caller. Direct-dial connections incur any applicable long-distance call charges. However, for the cost of a call, security can be significantly enhanced.

What are the implications for NCES of the ubiquitous accessibility of networks and the growing use of them by education-related agencies? The availability of universal network connections among NCES's trading partners nationwide provides tremendous potential and impetus for changes in the way data are collected and reported. This is not a new realization for the agency. In fact electronic exchanges have already been implemented in several areas. What this paper is pointing out is that now is the time to make that full commitment to use of electronic networks. There should no longer be a hesitancy to move forward as soon as possible with conversion of NCES data collections from paper to electronic.

NCES sponsored 30 automation feasibility site visits to state education agencies from 1992 through 1994. During these visits, numerous examples of states' early attempts at using floppy disks for submitting reports were found. Both visits that included higher education interviews found disk reporting being tried. Reactions were universally positive, and plans were in place for expansion of the process.

Software and Applications

Relational Databases

Whether in physical reality or in concept, the emergence of relational databases has changed how NCES can plan for the future. A relational database stores data in the form of tables. They are powerful in that they impose few assumptions about how the user is going to want to access or analyze the data. Consequently, many individuals can benefit from the same database by using it in many different ways. In contrast, a flat-file database is self-contained in a single file. Everything a user needs must be in that same file to be used together. Relational databases are ideal for large information systems. They are also ideal for systems that will be used by many individuals with contrasting information and analysis needs.

This database issue is important, because the future design of NCES information system needs to take into account that all the data that will be needed may not, probably will not, reside in one location—or even within the NCES LAN. In line with this, NCES is very unlikely to define a file structure that will become universal across all the data systems that contribute to the NCES information system. In this context, the relational database design allows for the accessing of information across files for analysis.

Electronic Data Interchange

Moving data directly from one computer to another is called electronic data interchange (EDI). EDI is used by businesses for items such as purchase orders and invoices. Within the past 5 years, EDI applications have been developed for student transcripts and college loan applications. In fact, NCES was a sponsor of the development of the SPEEDE/ExPRESS standards for student transcripts. SPEEDE/ExPRESS is an approved standard by the American National Standards Institute (ANSI). Several vendors offer software to perform the EDI exchanges of transcripts. The Far West Lab in San Francisco provided copies of their ExPRESS.cal application for the Migrant Education Program pilot.

EDI is basic to moving NCES from a forms-based paper system to a data file-based, electronic system. Some states that have already begun submissions of reports from districts to their state education agencies on disks use a different technique. These processes involve filling out what looks very much like the paper report forms on a spreadsheet or a word processing template, then making a copy to submit. EDI is the sending of a data record in a specific data format. The computer on each end of an EDI exchange can interpret the format and produce the types of reports on screens or paper that people are used to seeing.

Remember the last paper transcript you saw and compare that image to the format displayed in Figure 1.

This is an EDI record. The computer sending it and the computer reading it know exactly what each part means and how to interpret the contents. Each line is a "segment" containing information in one area. For example, the **SUM** line indicates 6 semester credits earned out of 6 attempted for all work taken at the sending school where 0 is the lowest possible grade average, 4 is the highest, 3.5 is the student's grade point average, and N means the grade point average cannot exceed 4. An entire transcript can be translated using these segments and their code tables.

You do not ever have to see this EDI language, because the computer translates everything into your local file format. When you see the information interpreted and printed as a transcript or displayed on a computer screen it looks no different than any other transcript.

In the absence of a national standard such as ANSI's SPEEDE/ExPRESS, commercial vendors would use their proprietary, and different, standards. Communications between vendors' systems would continue to be difficult.

Figure 1—Example of a SPEEDE/ExPRESS Electronic Record

```
ST*130874300021 N/L
BGN*00*87400021*900910*1530*ES N/L
ERP*DD*B48 N/L
REF*SY*123456789 N/L
DMG*D8*19790109*M*I*0*1US N/L
IND*US*FL N/L
N1*KR*Eastside Elementary School*77 *123456789101*9876 N/L
SUM*S*B*Y*6*6*6*0*4*3.5*N N/L
SES*198298*1**2*Fall Term*D8*19829824 N/L
SE*11*874300021 N/L
```

Productivity Software

Intelligent software applications that make work easier are emerging daily. The trend is for more of the work tasks performed to be automated. The benefits are not just for the worker who receives assistance with accuracy, finds the need to redo or recreate work less frequently, and is able to focus on more critical, clerical tasks. The benefits are also for the organization that receives data on the processes of the business and the work that is being accomplished. As the worker performs duties, the software does the work of keeping the records and producing the reports.

For NCES, the implication is that automated software applications can be developed that perform the technical aspects of reporting, look for and alert the users to data quality issues, and reduce the burden for those providing the data as well as for the NCES staff receiving the data.

Voice, Video, and Text Processing

An examination of the NCES data collection forms reveals that much of the information reported is textual. Software is available now to analyze the content of text, to search for key words, and to index topics. Voice recognition technology has advanced to the point where it is practical to translate speech into text. Imagine a performance report for Title I compensatory programs containing a voice message describing program implementation issues. Video is becoming a more common method for recording program delivery levels. Video is being analyzed for communications patterns. A combination of video and voice recognition could be used to create a text record of classroom activity, then to produce a content analysis.

Practical use of these technologies does not appear to be possible within the short term. The issues of interpretation and use would overwhelm staff who are already challenged by the quantity of data being collected. However, future visions and plans should recognize the potential for these types of data collections and analyses.

SUMMARY OF IMPLICATIONS FOR NCES

What does this all imply for NCES? NCES should position itself to ride the wave of automation in the nation. The trends described here are as follows:

1) Faster computers will allow NCES to expand the amount of data collected, analyzed, and reported while potentially reducing the time and burden imposed on clients and NCES staff. The burden imposed by the quantity of data collected will decrease as an issue over time. Burden will be a consequence more of the availability of data versus the need to collect unavailable data. Of the data that are a part of an existing automated system, the burden to pass them along to another agency for analysis lessens as computers become faster in processing large databases.

2) Increased storage capacity on computers will allow NCES to collect and maintain as much data as is reasonable to collect based upon the information needs of clients. The amount of data to be collected will not need to be limited by the problem of where to put them when they are received.

3) The universality of networks will allow NCES to collect data electronically, communicate to clients electronically, and make available its analyses and reports electronically. Not only will virtually all agencies have access to networks, they will be wanting to use them. There will be a demand from reporting agencies that NCES

accept all submissions electronically to avoid the burden of creating paper reports from local data files.

4) EDI standards and software will make electronic data exchanges over these networks efficient, effective, and affordable. EDI standards such as SPEEDE/ExPRESS may not become universal as the formats for maintaining data within agencies' databases. However, translations to EDI standards will become almost routine in order for agencies to exchange data files without rekeying information. In the short term, use of word processing templates and spreadsheets will begin the process of paperless reporting. NCES should continue to take an active role in the development of voluntary standards that facilitate electronic communications.

5) Relational database concepts will be applied to a distributed information system that will allow access to data across individual federal agencies' files. Where EDI standards provide a common language and process for exchange, database designs will allow for sharing or accessing of more complete data files by multiple agencies. For example, the Migrant Education Program in South Carolina envisions querying a data file in Georgia to locate the education records for arriving students. Then the Georgia schools will use SPEEDE/ExPRESS standards to send the students' records from their last school in Georgia to their new school in South Carolina.

6) Productivity software will automate information management tasks to the extent that staff will insist upon computer applications over any remaining manual processes. Software will continue to evolve to be more complex, more intelligent. Most of the tasks that do not require individual judgments will be handled by computers, with staff monitoring and intervening only when necessary.

7) NCES can achieve the benefits of an individual student-level database without the problems of creating a single one in NCES. The emerging networks and standards can create a national distributed information system. NCES would be able to query each state database to conduct analyses without having to maintain individual records centrally. The requirements for confidentiality can be maintained, and NCES would have access only to those data elements that are available to them by federal and state laws.

SPIN-OFF EFFECTS

The changes enabled by the advances in technology as described above do not come without their own spin-off effects. These are the indirect effects that occur as a consequence of a change.

Transfer of data processing responsibilities from a centralized data processing department/staff to the NCES staff or to the staff within other agencies is a major change. This transfer of responsibilities may also take the form of moving tasks from a few key staff members to a larger set of workers. As productivity software is installed, as networks make direct connections between agencies, as agency staff perform the actual data management tasks, the need for an external service group traditionally called the data processing department changes. This has benefits when staff are no longer waiting for their work to move up the priority list.

Data are on your own computer, available when you need them. This has a downside when your staff must be retrained to perform new duties.

The role of the traditional data processing department shifts from one of actually doing the processing to one of supporting those who are. Programmers and systems professionals who are grounded in mainframe computer operations can have a difficult adjustment to the very different skills required in a distributed information systems environment. Data processing professionals will be called upon to support others and their applications.

The quantity of data will increase, especially as nonaggregated data are reported. More data and more analyses will put pressure upon staff to monitor and assure the quality of statistics and the reliability of analyses. Quality assurance procedures will need to be adjusted accordingly. Today, NCES calculates and issues official statistics on the nation's schools. With a data warehouse providing access to many researchers and interested organizations, almost anyone can calculate his or her own versions of those statistics. This would lead to a healthy debate as alternative analyses and perspectives are examined. This can also lead to the necessity for NCES to defend their formulas and calculations. Some form of quality check will be needed to respond to the alternative statistics offered by individuals and organizations. All of these will not follow the same rigorous standards NCES staff will follow when producing statistics.

IMPACT OF TECHNOLOGY CHANGES ON HUMAN RESOURCES

What businesses have discovered and learned to plan for is the impact of changes in technology on their people. Hardware and software costs are usually less than the associated costs for training and supporting the users. Within education organizations, the impact could be even greater. Staff development has historically received low priority—even for activities that are clearly directly related to the primary learning focus of the organization. Much less emphasis has been given to technology- or data-related issues.

Beyond retraining individuals and modifying hiring requirements and practices, organizations must restructure their staffing charts to reflect changes in the activities of staff. For example, state education agencies are already changing formerly secretarial positions into software applications support and training positions. As managers do more of their own word processing, there is less to type, and other traditional secretarial tasks also decline.

For NCES, planning must recognize the changes that will be imposed upon other agencies who must adjust to more automated processes. NCES will need to consider its role in retraining state and local staff. Development of training materials, sponsorship of workshops, and other support should be considered. NCES and its trading partners will be revising their job descriptions and the qualifications sought for new staff. Promotion and assignment decisions will reflect more of the technology-related skills necessary to implement and maintain the automated systems discussed here.

A VISION OF FUTURE AUTOMATED INFORMATION ACCESS

NCES will create a vision for future data collection and reporting. With the technology advances described in this paper, the following aspects of a vision seem reasonable:

- There will not be reports to fill out and submit. The concept of a report will change from being a document that someone fills out by collecting, calculating, and entering information. A report will become an analysis created from data sources available within an organization's information system.
- Most of the surveys and data collections that occur now will disappear. The concept of a survey or data collection as a specific request made for information on a report form will change. The individual needing data will go directly to a data file and read/copy what is desired. The concept that a survey or data collection occurs at a given time will shift. Data can be harvested from data files as needed, multiple times during a year. As an alternative, a reporting agency can upload (submit) their data as they become available.
- Almost all data about education will come directly from databases that are built as a natural part of conducting the business of an education agency. As more work is automated using productivity software, data documenting that work will be maintained as part of the software's task. Grades within automated grade books, records of transcripts sent to colleges, numbers of free meals served, and so on, will be recorded as these actions occur. When the data are needed, the data files will be read directly.
- When a new mandate for data collection and reporting arises, existing data sources will satisfy most of the requirements. New mandates for information will be checked against existing data sources. Only those elements that are not already available will need to be added to the information system.
- School personnel and education agency and staff will not think of the paperwork burden imposed by other agencies, because most of it will be transparent—accomplished as routine within their own automated management systems. Instead, considerable thought will be devoted to keeping information systems compatible, linked electronically, and current.
- Data will be collected and entered into these management systems because they are useful to the schools and education agencies. The best quality control is achieved when the persons responsible for the data depend upon the data for their own purposes. When the data have meaning, the individuals responsible for the data know when they are accurate and complete. Burden will not be a major issue, because the data are useful to those producing them.
- When the educators, news media, researchers, parents, and others have an information need, they will access data directly through an electronic network, in their own offices or homes, and create just the reports they need. The concept of huge volumes of statistical reports will change. The statistics will exist in data warehouses rather than on paper. Some statistics may not even exist until they are requested. Many more statistics can be produced than would be in a printed volume. Audiences can access statistics or in some cases the data used to calculate those statistics.

- Confidentiality will be maintained within the automated systems, allowing access to those with clearance and denying it to others. Directories and certification processes will determine an individual's access to data.
- A common data dictionary will define data elements and statistics along with the periodicity of their collection. Agencies will voluntarily use common data dictionary entries to ease the burden of translation when information is exchanged.
- Electronic networks will connect agencies, so data can be harvested from databases according to the periodicity specifications. Agencies will be able to read data directly from each other rather than having to make a request and await a reply. Data within each agency's information system will be categorized as public, restricted, or confidential to ensure that confidentiality rights are protected.
- The system will be voluntary, and compliance will be almost universal. Compliance will come from a common understanding of the benefits. Some entities will choose not to automate, and others will have local laws limiting participation.
- Cost savings will offset expenses, and the savings in personnel time will refocus resources on the primary mission of the educational agencies. Teachers will have more time to teach, librarians will have more time to manage their collections, financial aid officers will have more time to counsel students, etc.
- The components of the system will develop over time, joining together as they become available. Every agency will not participate from day 1. A paper system will be needed for some. Over time, the vision will become more universal.
- NCES will enable the system to develop by setting national standards and encouraging states to follow their example. The role of NCES will be key. As a facilitator of standards and a collector of data at the national level, NCES will be a model, a sponsor, and a participant.
- Reports will be printed by users as they are needed; many will be read on a monitor and no paper will be used. The concept of printing and disseminating a report will change. Most reports will be placed within the data warehouse and audiences will access the parts they need. Printing can occur at the reader's location rather than at the Government Printing Office. Printing would be at the reader's expense.
- The quality of education data will improve dramatically as use of the data motivates everyone toward accuracy, and the source of data becomes the management system that educators depend upon for their own work and productivity. As the data are used by more individuals and for more purposes, the benefits of accuracy and the risks of poor data increase.
- The ultimate purpose for collecting, analyzing, and reporting education data is to improve learning. With an open information system informing decision making, improvements in the quality of instruction and the management of education agencies will occur at a faster pace than ever before.

To achieve this vision, NCES will need to employ technology effectively. A major part of its planning must include a data warehouse or an alternative that achieves the same level of access to its information resources.

THE NCES DATA WAREHOUSE

Currently, access to NCES data and publications takes the form of printed documents distributed through a dissemination process involving mailing lists and orders through the Government Printing Office. Recently, some NCES publications have been placed on an Internet World Wide Web page for access. Access in the future should have many options from print to electronic files.

The technology trend and advances described in this paper support a direction already evident in NCES's planning—to develop a data warehouse. A data warehouse is simply a location where someone can access information electronically. As with many terms in the technology arena, there are differences in the characteristics people attribute to a data warehouse. A major attribute that varies across users of the concept is the level of aggregation for the data provided. To some a data warehouse is like a library containing books with statistics and analyses already accomplished and described. To others, a data warehouse contains an organization's raw data—available for analysis. For NCES, both are appropriate. With very few exceptions, NCES's data are public, as are any documents produced. Therefore, protecting the confidentiality of data or limiting the distribution of reports is seldom an issue.

NCES is on target with its current effort to build a user-friendly interface with its data warehouse. The key to widespread use for any computer application is utility and ease of use. NCES's concept is to give users the ability to search files for the data or other information they are seeking, then to download them as desired. The contrast with this and the current printing of large paper volumes called digests of education statistics is mainly with the ability of the user to find what is sought online rather than to find a printed volume and look up the statistics. An added bonus for users will be the ability to create tables and reports containing the information in which they are interested, rather than being limited to the manner in which data have been presented on the printed page.

The data warehouse can also function as a receiving point for data. Submissions by states can be uploaded to the data warehouse as soon as they are ready. This method can also be integrated with the harvesting concept. Both can operate within the information system.

Of course, the data warehouse concept should not stop with NCES. In fact, at least one state, Hawaii, has a functional data warehouse now, and others have them in the planning stage. The description that follows considers the benefits of a collection of data warehouses that are connected by networks and common EDI standards.

In this possible model, there would be multiple data warehouses containing in the aggregate all of the important and useful education data from across the nation. NCES would have one. Many individual states would have one each. Some states might join together to share a common data warehouse. Some states might use a commercial service. Within some states, there might be regional centers that provide this function. Some districts may be large enough to justify operating their own. Even some schools, especially private schools, may want to establish their own. The fact is that the number and nature of the individual data warehouses and who is participating in each is not consequential. What is important is that they all use certain standards for EDI. They might also all use common database structures or formats to allow direct access to selected files by other organizations.

In the diagrams in Attachment A, NCES is shown as building and maintaining a central directory of agencies. This directory would build upon the Common Core of Data directory information currently collected. In addition to current data elements, this electronic directory would contain each agency's network address, contact persons, access information, and other usage parameters. The directory could be updatable directly by each agency. Thus, it would become a self-maintained directory.

The collection of data warehouses would be a distributed information system to the extent that common standards are used to store and access the contents of each. The contents accessible this way would be restricted to those data elements that each agency is authorized to provide to other agencies. This set of data is called the Confidential Data File. Contents would include items such as individual student and staff demographics, immunization data, course and grade data, assessment results, and program membership data.

A second data file within each data warehouse would be called the Public Data File. The contents of this file would be available to anyone. This would include such items as aggregate demographic statistics, enrollment statistics, financial data, assessment reports, and campus descriptions.

Behind these two files that are accessible to persons outside the agency would be the source data files. These source data files would be the master copies of data and would contain all data elements. These files would be secure, and users of the data would access copies of these files.

National education data and publications would reside in the NCES data warehouse. Communications between data warehouses or with individuals would be through the Internet, VANs, or direct dial as established by each agency.

How would the existence of these data warehouses affect NCES's data collection processes? Instead of sending out surveys to be completed or other forms-based data collections, NCES could connect to each data server for each data warehouse and download the information needed. The timing of these downloads would have to be known by all. Each data server should also contain an indicator of the status of the data for download by NCES. Each agency would be left an electronic receipt for their data.

In order for the data warehouse network to function, there must be national standards for data definitions and formulas. This is equivalent to a common data dictionary. However, even without a common data dictionary, participants in the distributed information system can communicate by translating their local data to a common standard such as SPEEDE/ExPRESS.

In Attachment B, the relationships among the levels of education agencies are described. The data within each level's information system are shown as being for internal use only, or as being shared with other levels. For either direct reading or harvesting of data to function, these relationships must be clarified and the data elements that fall within each category must be identified.

Timeframe for the Vision

Portions of the vision are in place now. Some states and some NCES activities are following, or more appropriately, leading the vision. The technology required for this vision to be fully implemented is already available. The hardware and network components are the most advanced. The productivity software will continue to be developed as agencies call for it to advance. It is reasonable for NCES to target converting all of its data collections to EDI by the year 2000. Activities may need to provide for paper submissions as an alternative for some.

The transition of NCES to automated data collection and a data warehouse is an ongoing, developmental process. There is not a turn-key system that can be purchased and installed.

Assumptions for Planning the Future Systems of NCES

The previous discussion of the advances and trends in technology points toward a set of assumptions that NCES should consider in planning its future information systems.

1) NCES can expand the amount of data collected, processed, and reported using faster computers. The time and burden imposed on clients and NCES staff will be less because of this processing efficiency.

2) NCES can collect and maintain as much data as is reasonable based upon the information needs of clients. Increased storage capacity on computers will allow reporting agencies and NCES to handle significantly larger data sets.

3) NCES can collect data electronically, communicate to clients electronically, and make available its analyses and reports electronically using national networks. Current forms-dependent data collection systems can be replaced with EDI-based systems with the expectation that reporting agencies can comply and participate.

4) Electronic data exchanges over these networks will be efficient, effective, and affordable. EDI standards and software will make these exchanges practical for agencies.

5) The change to EDI and other automated systems will require significant retraining of staff at all levels.

6) Allowing direct access to information in a data warehouse will increase the use of NCES information.

Process for Exchanging Data Among Education Agencies

THE CHALLENGE: To establish a process that connects schools, districts, states, and NCES for the electronic exchange of education data.

This challenge is great because records across schools, districts, and states exist in various formats and on a variety of computer systems. In addition, there is no funding available, even if agreement were to be achieved on the design, to build a single system. The ultimate solution must be an open-systems design that accommodates state, regional, local, and vendor solutions.

National Center for Education Statistics Data Warehouse

This data warehouse functions to provide all audiences access to the data collected, the analyses conducted, and the publications produced by NCES. Electronic access through telecommunication networks provides immediate access without the necessity for NCES to print and mail copies.

WORKING ASSUMPTIONS:

1. **States will adopt different models to achieve the common goal of electronic data exchange.**

2. **Vendors will offer commercial solutions.**

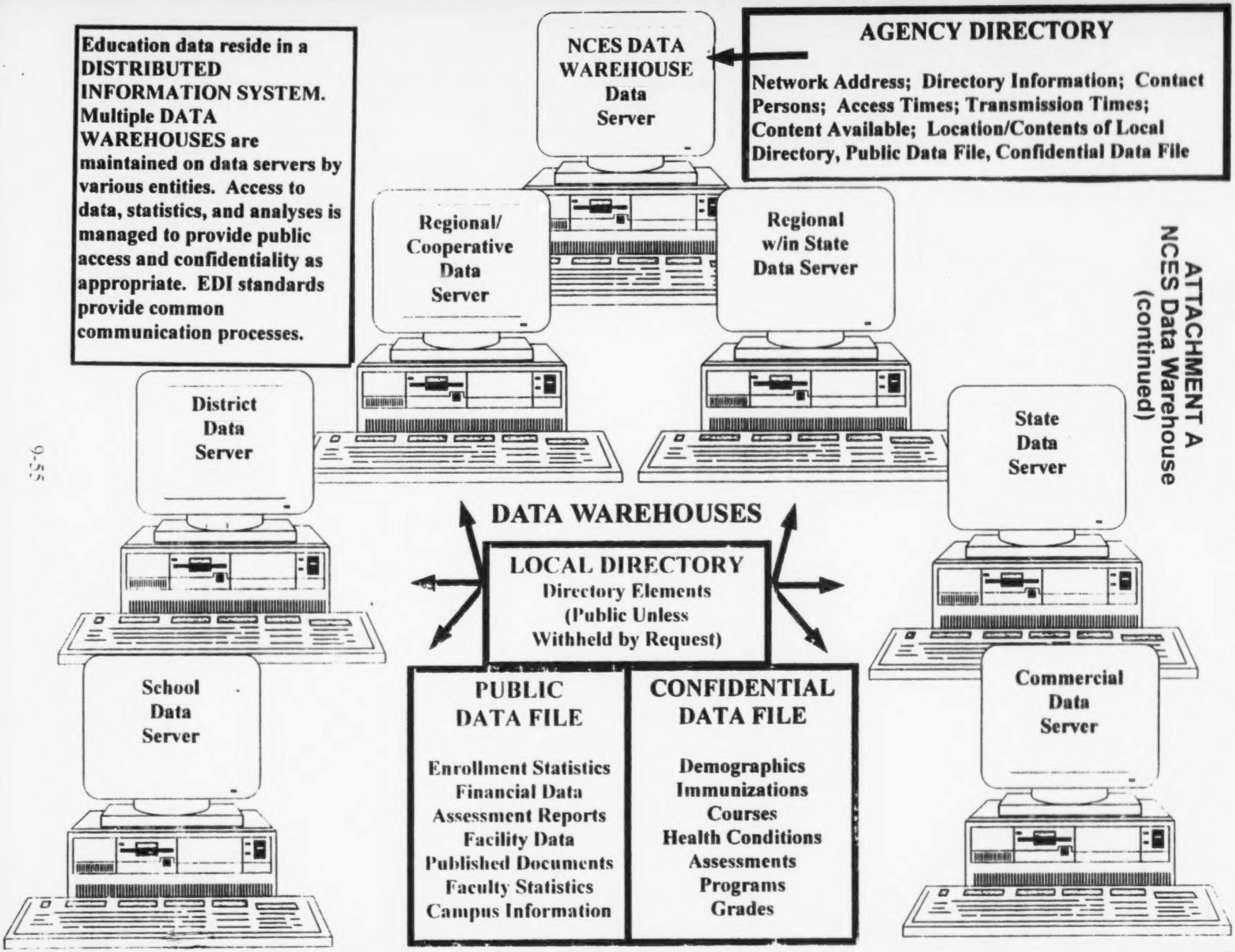
3. **The role of the federal government will be to facilitate the development of standards and to use those standards for data collection. The federal government will not fund and build a system.**

4. **Participating schools, districts, and states will maintain control over access to their records.**

5. **Participating schools, districts, and states will determine the data they exchange.**

ATTACHMENT A
NCES Data Warehouse

ATTACHMENT A
NCES Data Warehouse
(continued)



Education data reside in a DISTRIBUTED INFORMATION SYSTEM. Multiple DATA WAREHOUSES are maintained on data servers by various entities. Access to data, statistics, and analyses is managed to provide public access and confidentiality as appropriate. EDI standards provide common communication processes.

AGENCY DIRECTORY

Network Address; Directory Information; Contact Persons; Access Times; Transmission Times; Content Available; Location/Contents of Local Directory, Public Data File, Confidential Data File

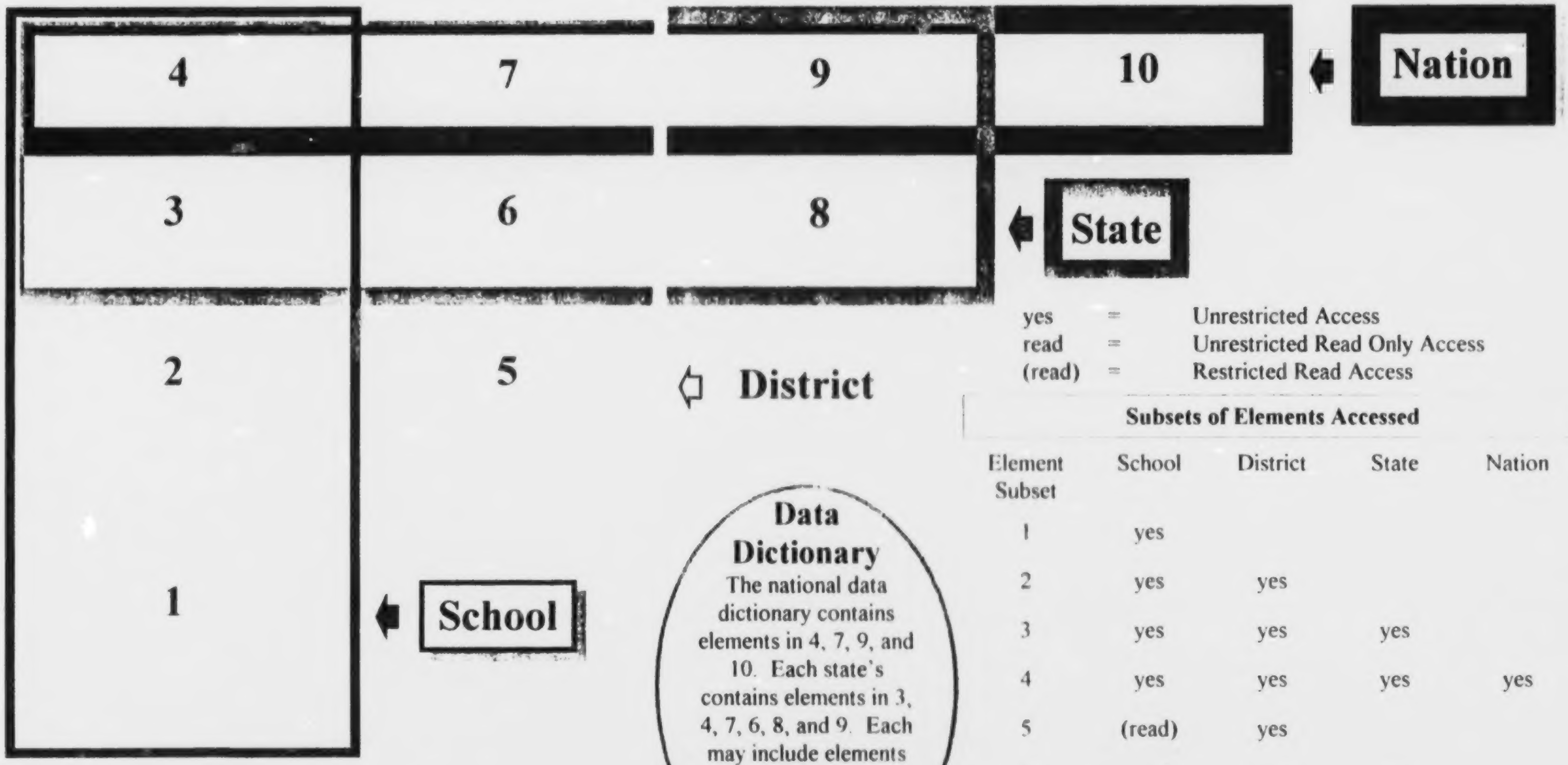
NCES DATA WAREHOUSE Data Server
Regional/ Cooperative Data Server
Regional w/in State Data Server
District Data Server
State Data Server
School Data Server
Commercial Data Server

DATA WAREHOUSES

LOCAL DIRECTORY
Directory Elements
(Public Unless Withheld by Request)

PUBLIC DATA FILE	CONFIDENTIAL DATA FILE
Enrollment Statistics Financial Data Assessment Reports Facility Data Published Documents Faculty Statistics Campus Information	Demographics Immunizations Courses Health Conditions Assessments Programs Grades

Automated Information Coordination
Rectangular Venn Diagram for Sharing Data Across Agencies



yes = Unrestricted Access
read = Unrestricted Read Only Access
(read) = Restricted Read Access

Data Dictionary

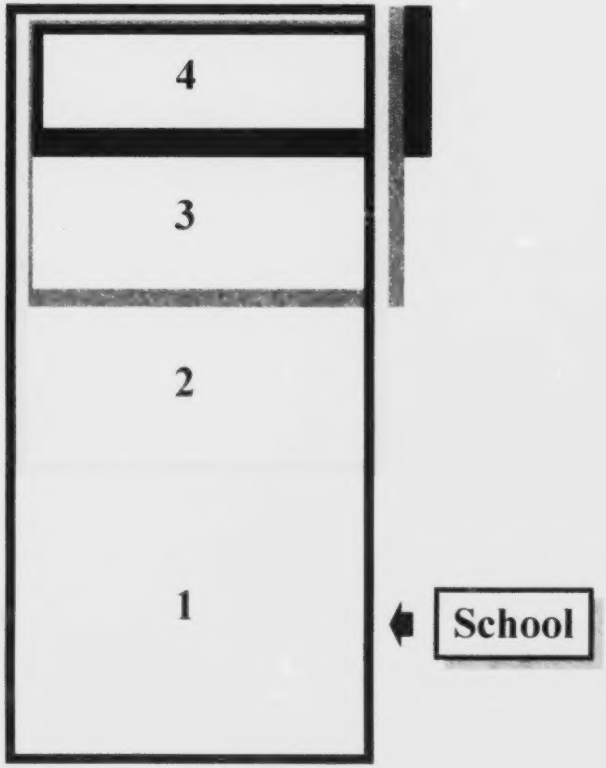
The national data dictionary contains elements in 4, 7, 9, and 10. Each state's contains elements in 3, 4, 7, 6, 8, and 9. Each may include elements in other subsets to maintain comparability.

Subsets of Elements Accessed

Element Subset	School	District	State	Nation
1	yes			
2	yes	yes		
3	yes	yes	yes	
4	yes	yes	yes	yes
5	(read)	yes		
6	read	yes	yes	
7	read	yes	yes	yes
8	(read)	(read)	yes	
9	read	read	yes	yes
10	read	read	read	yes

Example A school maintains element subsets 1-4 Elements in 1 are used only by the school, whereas elements in 4 are shared by the school, district, state, and national agencies

School-level data are characterized by being more individual to students, more anecdotal, more personal, and more instructional.

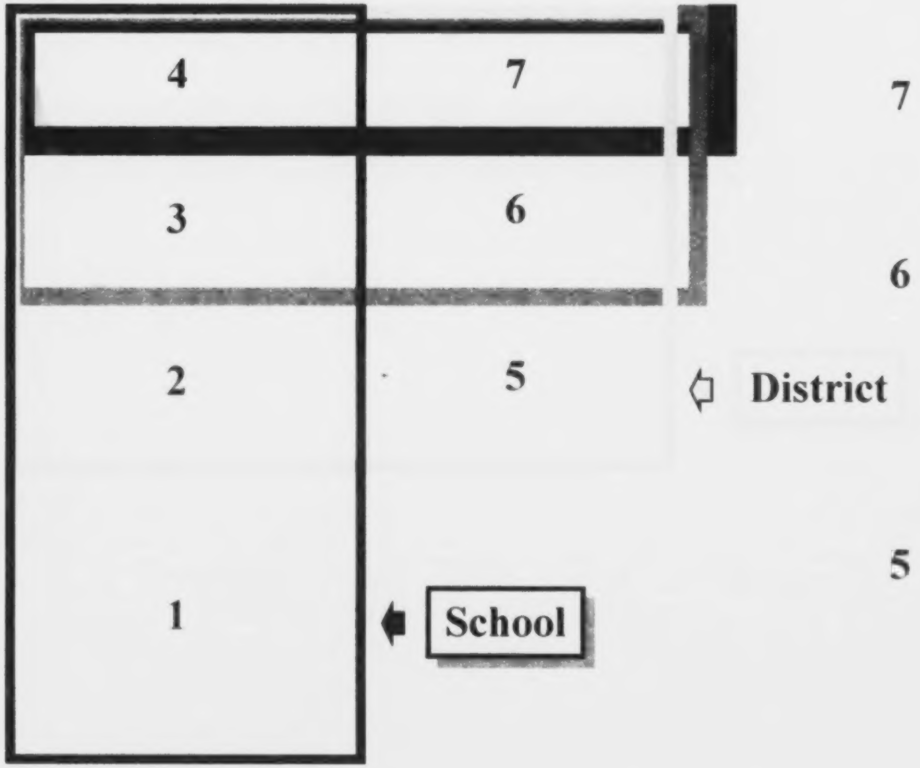


4	School-level data typically exchanged with national level: Gender, ethnicity membership counts School address, phone Grade levels served
3	School-level data typically exchanged with state level: Attendance for funding formulas Gender, ethnicity membership counts Title I participant counts
2	School-level data typically exchanged with district level: Attendance Gender, ethnicity membership counts Teaching assignments
1	School-level data not typically exchanged with other levels: Book club orders and payments Homework and daily grades Emergency contact numbers and names

School data exchanges above the district level are typically performed by the district.

ATTACHMENT
B
(continued)

District-level data are characterized by more aggregated statistics.



7 District-level data typically exchanged with national level:

- Gender, ethnicity membership counts
- Title I participant counts
- District address, phone

6 District-level data typically exchanged with state level:

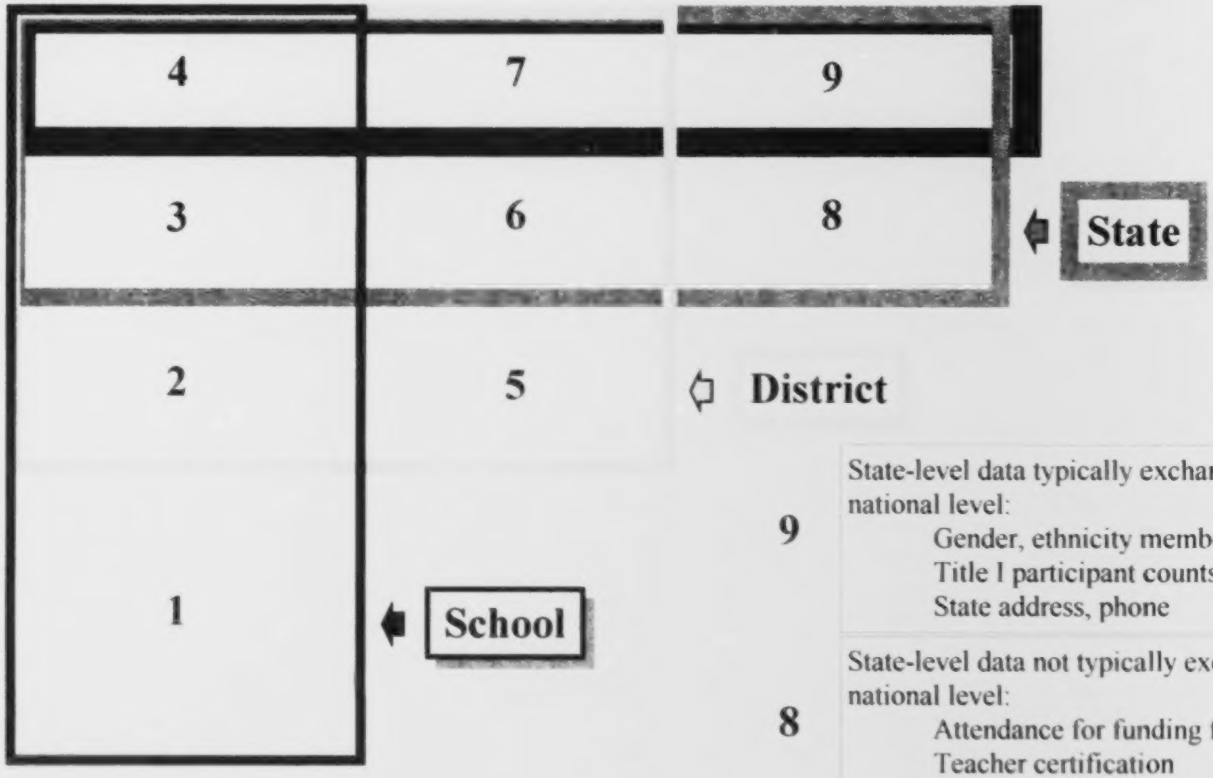
- Attendance for funding formulas
- Gender, ethnicity membership counts
- Teacher certification

5 District-level data not typically exchanged with other levels:

- Statistics on test scores
- Attendance area statistics
- Committee membership

District data exchanges above the state level are often performed by the state.

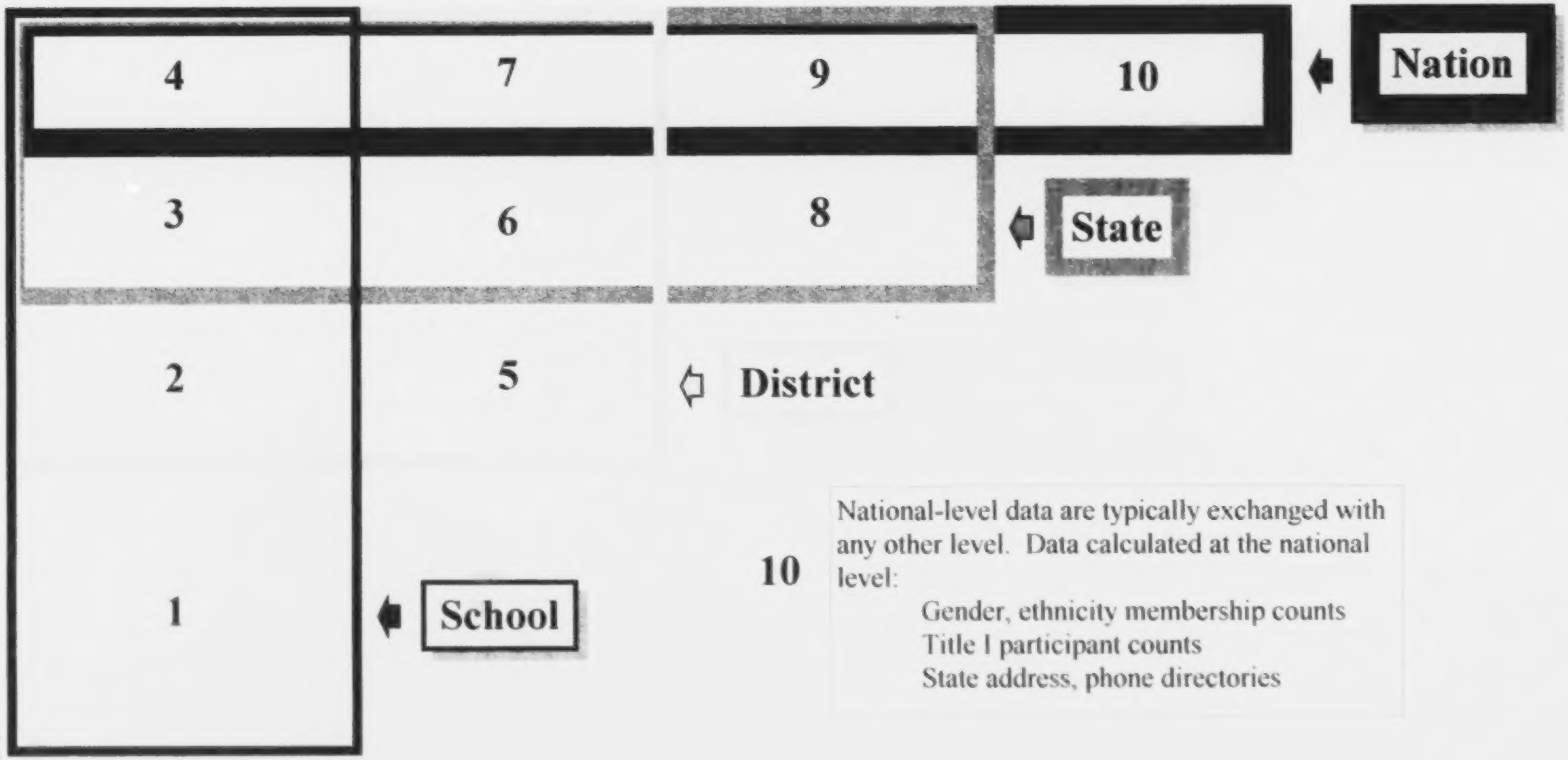
State-level data are characterized by mostly aggregated statistics, except for states with student-level databases.



9 State-level data typically exchanged with national level:
- Gender, ethnicity membership counts
- Title I participant counts
- State address, phone

8 State-level data not typically exchanged with national level:
- Attendance for funding formulas
- Teacher certification
- State assessment statistics

National-level data are characterized by almost exclusively aggregated statistics.


4
7
9
10
Nation
3
6
8
State
2
5
District
1
School

10 National-level data are typically exchanged with any other level. Data calculated at the national level:

- Gender, ethnicity membership counts
- Title I participant counts
- State address, phone directories

Many Tenets of the Current Information Systems Will Change

The periodicity of collecting data will evolve from dates when forms are available and completion begins, to timeframes when files are built and extracts occur. Due dates will become extract dates, representing when data files will be read.

Data burden will shift from a term depicting a situation where work stops to document and report, to one where work is automated for efficiency and data for reporting are a byproduct of work activities. Data burden may become a term representing the overwhelming amount of information to read and interpret that is available on a given topic.

The producers of data, the entities being monitored or reported, will not have complete control over the information produced and known about them. Access to raw data will delegate to others the ability to generate statistics not ever seen by the target of the analysis. This loss of control by educators at all levels could slow the progress toward fully automated information systems unless groundwork is laid along the way.

Physical separations will be established between systems. Data warehouses will be created to hold the data that are accessible by a credentialed set of users, but the original management information of an organization will be more tightly controlled on isolated computer systems with fewer if any access links outside the physical facility of the organization. This means that levels of access will be established within this firewalled city of information. The sanctum sanctorum of an organization's data will be the original secure source files containing the most detail and most confidential elements. Confidential working extracts will be created and posted for certified users with a need to know. Multiple extracts will be created with the set of data authorized for a set of users. These extracts will be loaded onto separate servers without access beyond the organization and with only one-way access from the secure files. Limited access extract files will then be created, again directly from the secure source files. These files will be available on more universally accessible file servers, access to which is allowed for certified users from certified locations. The fourth and final set of extract files will contain public information placed upon a public access server in a read-only mode.

OUR CONCEPT OF DATA WILL EXPAND

Our concept of data has begun to expand as storage and processing capacities increase. Images, video, and voice have all claimed places within our automated data systems. How we will analyze and use them is expanding rapidly already. Content analyses, image scanning, voice-to-text translators, image-to-text translators are all becoming more sophisticated and allow analysis of now seemingly insurmountable amounts of data within a reasonable time in an automated fashion. Classroom observations recorded on videotape for example could be analyzed and coded with software programs designed to detect who is speaking, topics, movement, involvement of individuals, and even performance feedback actions.

CRITERIA FOR JUDGING THE FUTURE SYSTEM

In December 1985, Hall, Jaeger, Kearney, and Wiley prepared a report entitled *Alternatives for a National Data System on Elementary and Secondary Education*. Within that report, they proposed a set of criteria to be used for judging a national education information system. These criteria are revisited here along with a comment related to the characteristics of a system that would be consistent with the vision described above.

1) Comprehensiveness—the system must have a database capable of providing information on all pertinent aspects of elementary and secondary schooling, including the school setting, the schooling process itself, and the outcomes of schooling.

   Because the vision foresees inclusion of all data that are produced as a product of the conduct of the regular business of education agencies, it should be comprehensive. A limitation would be that data on the school setting and the schooling process would be available only to the extent that automated systems are used which would relate to school setting and processes. Is it the role of NCES to collect and process data? Certainly the vision sees an information system that includes a much more comprehensive database than is currently compiled.

2) Integration—the elements, files, and records in the database must be linked; all data sets must be capable of being related to one another.

   A relational database design would facilitate this. A common data dictionary would be necessary along with definitions and formulas for all calculated statistics. This criterion would require that links already exist at the local, district, and state levels.

3) Micro Record Format—all data must be collected and stored in micro record format, with a micro record being defined as a datum on an individual person or an individual entity.

   This is problematic at the national level. Despite the increased storage capacity and speeds of computers, a data file of over 50 million student records would be cumbersome. This criterion can be met if we accept the idea that there will be such a micro record for every individual at some level of the distributed information system. Some individuals records will reside only in a school's database, others at a district level. A few states have individual records systems for students; many more have them for staff. There is no mandate or plan for NCES to collect personally identifiable records at the national level.

   However, with the ability to harvest data or conduct analyses on data distributed across multiple data servers, the functional intent of this criterion could be met.

4) Representativeness—in addition to being nationally representative, the information in the database must be representative of each of the 50 states, as well as representative of other important variables such as sex, racial-ethnic composition, urbanization, and so on.

   This criterion would be met with the participation of all states. The content collected by current Common Core of Data surveys addresses the intent of this criterion. The collection and storage of disaggregated data, individual student records, and generally more detailed data provide the opportunity for post hoc analyses that consider additional variables.

5) Accuracy—all data must be verifiably accurate; they must be subjected to rigorous quality control procedures including audits, reinterviews as a routine part of data collection, controls on data entry and data processing, consistency and completeness edits, and regular and routine calculations of measures of variance.

The emergence of large, central databases into which data are reported for state information systems has popularized the term "desk audit" to represent quality checks that are performed on data that have been reported. An individual sitting at a personal computer on a desktop can run verification and audit software. These checks look for data out of normal ranges, illegal codes, missing data, etc. NCES or independent groups would be able to perform reasonableness audits or even follow up with source rechecks if data are provided in a data warehouse.

Productivity software used would contain validation checks as data are entered (the level where errors are most commonly created and where they are most easily resolved). EDI software contains validation checks for data sent and received. However, errors that fall within the normal range of data can typically be found only by the provider of the data.

6) Comparability—data from different jurisdictions must reflect the same concepts and definitions; common units of reporting and common definitions are necessary precursors of useful data aggregations.

NCES has traditionally provided clear definitions and formulas for the aggregated statistics it collects. How faithfully the data providers follow these standards varies. Automating the collections will not solve any current problems. However, adoptions of common data dictionaries and use of common software applications can emphasize the definitions that are to be followed.

7) Timeliness—in general, data must be limited to that which can be collected, stored, and analyzed within three months and reported to policy makers within the year.

The vision uses technology to address this criterion directly. Taking data from existing systems, electronically exchanging them, and providing the capability for faster analyses using large data sets all contribute to timeliness. Reporting to policy makers can be improved with electronic availability from the data warehouse. One concept with the data warehouse is that states could post their data as soon as it is available, then they would be accessible by others immediately. The existence of a data warehouse can shift the burden to the users to know when data are available.

8) Privacy and Security—because some of the elements, records, and files contain information about individuals (e.g., personal identifiers necessary for longitudinal studies), strict confidentiality and security measures must be in force.

Confidentiality and security challenges exist with paper systems. They receive greater attention with automated systems because of some highly publicized events and the very real risk of hackers. Electronic systems allow for very elaborate security processes. Even these are not failproof. However, the required sophistication of a successful hacker can be pushed higher and higher, and automated systems can document access. Many individuals believe that the security of electronic systems is superior to that of paper systems.

9) Processing and Analysis—a specific schema must be available for processing the micro records in a manner designed to optimize the analytic capacity of the system.

   The increased speed and storage capacities of computers contribute to this. The ability to analyze the larger data sets has improved considerably since 1985. Analysis software and the emergence of relational databases have boosted the capacity of researchers to perform analyses.

10) Information Flows—the system must be capable of screening and matching its reports to meet the particular needs of users; a wide array of reporting formats and access mechanisms must be available to serve the different users; specific priorities must be set for meeting the different timelines imposed by the needs of the users.

    This is an excellent example of a criterion that is well served by technology. With a data warehouse, users will be able to search indexes as well as text to find information matching their needs. Reporting formats increase with the addition of screen views, downloads, and user queries to produce just the statistics desired. With electronic access, users can get the information they need when they need it. The only constraint is that the data must be collected and already captured by the information system.

11) Costs of Transmission/Access—a pattern of shared user costs should characterize the system; rather than rely exclusively on federal support for transmitting information to users and/or providing them access to information, a national educational data system should also draw from a program of user fees and thereby increase its capacity to serve the differing needs of its users; equally important, transmission/access modes should incorporate the latest developments in electronic communications technology.

    The user pays the cost when connected through an electronic network. Whoever is connecting pays the transmission fees. The costs for establishing and maintaining the data warehouse would not be easily shared with the users. The cost of that type of billing might exceed the actual fees recovered.

The conclusion of the authors was that the only criterion met by the NCES system of 1985 was Privacy and Security. Interestingly, this is the one that could be the most controversial with an electronic system. For the other ten criteria, an automated system using a data warehouse concept has the potential for significant improvements.

CONCLUSION

Ensuring that NCES's data collection, storage, analysis, and reporting processes take full advantage of technology will be a process, not an event. This transition will require considerable training and support for both NCES staff and the staff of its data providers. When evaluated against the criteria described in 1985, the vision of the future as described here would be a significant improvement over past and current systems.

Discussant Comments

BARBARA S. CLEMENTS

These comments address issues raised in two papers: *Administrative Record Opportunities* by Fritz Scheuren and *New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education* by Glynn Ligon. Both papers describe important issues that must be considered by the National Center for Education Statistics (NCES) as it seeks to make its data collection activities more efficient and as it responds to technology changes occurring in the sites where the data originate. In these comments, I provide some background comments, and then react to the papers from two perspectives: the user perspective and the provider perspective.

Administrative records exist in all schools, districts, and state education agencies in a vast array of formats and with a variety of contents. While many schools, districts, and state education agencies may have some data automated, most are still heavily reliant upon paper records. Two examples illustrate this point.

About 10 years ago, when Texas was implementing a Career Ladder, a teacher from a tiny district called to see about getting evaluated for the Career Ladder. In the course of the conversation, she was asked where her personnel records were kept. She thought for a minute, and then said that she believed they were in a shoe box under her bed. Eight years ago, when I moved to Washington, D.C., I went to my son's school to get a copy of his high school transcript. I was given a photocopy of a paper document that had computer labels pasted on it. It was obvious that some parts of his student record were computerized, but the paper document was still used to compile his course data. According to my school contacts, these two examples illustrate the lack of technological sophistication with administrative records that still exists today at the school and school district level. I have heard of very few places in elementary and secondary education where there is a fully automated administrative records system that can handle the types of electronic exchanges and sophisticated analyses that are technologically possible today.

How data are used at the local and state levels is important when considering data quality. My sense is that in most schools and districts, most data are recorded because someone thinks they should or because someone requires it, such as the state legislature or the federal government. Few state or local education agency staff members have the time or opportunity to think about how data can be used to assist in providing quality instruction to children, the primary goal of the education system. Since the data have "little utility," there is no impetus to ensure comparability or timely updating. If NCES is to get useful data from state and local administrative records, it must develop ways to encourage and help data providers to collect and provide comparable, complete, and timely data.

Data User Perspective

As a data user, I have several comments about the papers. The Ligon paper describes the design for an automated administrative records system that can provide data access and give flexibility for data analysis to all levels of the education system. The Scheuren paper describes what valuable information is available when administrative records can be collected. Timely data availability is an important benefit both authors describe, and it relates to the ease with which electronic administrative data can be transmitted to different levels of the education system.

Current lag time in getting data from NCES from the Common Core of Data and other surveys has been frustrating for many data users. The work that can be done by NCES to streamline data editing routines and speed up reporting and data tape availability is essential. An electronic system such as the Ligon paper describes can allow data to be submitted from original sources with no rekeying needed; thus, the errors in the data should be minimized, and this should speed up the process of making data available to users. Such a system requires preparation at all levels of the data system; therefore, it is important for NCES to be ready to accept electronic data and process them quickly and efficiently.

Both papers indicate that moving to electronic submission of administrative records can provide more comprehensive sets of data with which to work at NCES. Each time NCES asks for new data elements to be added to paper survey documents, there are state education agency staff members who complain about the burden of adding those data elements to their own collections and the lag time that is needed to get data from all sources. If states have access to electronic administrative records, it should be easier for them to get additional data elements if deemed necessary and provide them to NCES. This would make the data sets more complete and better able to respond to both policy questions that arise in Washington, D.C. and to questions asked by other NCES data users. This is another good reason for NCES to continue working with state and local education agencies to design automated administrative records systems with electronic transmission.

To me, the most important thing that should be stressed in the discussion about administrative records is the need for comparability in what is collected and provided to the different levels of the education system. NCES has been working for years with state and local education agency staff to build a consensus on how the data should be collected and reported to ensure comparability. This is stressed in the Ligon paper, but not in the Scheuren paper. Although all of the data maintained in administrative records at all levels of the education system need not be exactly the same, the portions that are reported up from the lowest levels must be comparable, or at least able to be crosswalked, in order for the data to be useful. Therefore, as a user, I believe it is important for NCES to continue efforts to promote comparability and standardization of those data elements that are essential for national data collection.

The Scheuren paper suggests that administrative data be used to track changes over time. I believe there is a real need to look at changes in student population, effects of participation in programs based on new federal or state policies, and other educational issues that can help decision makers in planning for school improvement. Besides tracking changes, NCES needs to explore ways of identifying effective programs through regular data collection activities, so that case studies or further research can be done, not perhaps by NCES, but by others within the Department of Education, such as the OERI institutes.

Data Provider Perspective

There are several comments I would like to make from a data provider perspective. The work that NCES has supported related to providing tools to make the collection and transmission of administrative records easier are to be applauded. Burden is one of the most frequent complaints of state and local education agencies. State and local education agencies are looking for models of electronic data sharing that would be relatively easy to implement in technologically unsophisticated sites, and particularly ones that take into consideration existing equipment and planning for a system that can be implemented over time as funds become available. Such models would help state and local education agencies reduce their reporting burden and move toward providing more timely data. NCES has done some work to provide models for how data can be maintained, transmitted electronically, and used more effectively. The work NCES sponsors on confidentiality is extraordinarily important for all levels of the system. These activities have a great potential for payoff, and should continue.

Several areas still need the attention of NCES. First, NCES should look at all of the areas in its surveys where administrative records could provide essential data such as years of teaching experience, age, and so on, and plan to collect data in this way from schools, districts, or state education agencies to reduce the individual burden of individuals such as teachers who complete the surveys. To help promote comparability, stress should be placed on standardizing those data elements that will help data providers adjust their systems (or purchase appropriate systems) to meet future data reporting needs. As my data provider friends say, "Just tell me what you want and how you want it, and we will make it happen."

Second, many data providers need help with training on how to collect, report, and use data. At present, NCES provides a valuable service through the Fellows Program. Many state and local data providers would appreciate having models for how data can be presented more effectively for decision makers. For instance, videotapes are considered extremely useful by data providers because they can go back and review them when needed. Moreover, state data providers need help in training data providers from the local levels. Training is essential to getting comparable, complete, and timely data. NCES should place an even stronger focus on what they can do in this area.

And, finally, NCES should lead discussions with the health and human service areas about data sharing for the benefit of students. In education, we are constrained (and helped) by the Family Rights and Privacy Act (FERPA), and the other areas also have their professional ethics or other types of restrictions on usage. Currently, an important trend is on providing services to students through the schools. We are also encouraging teachers to make better use of student data when planning learning activities. NCES can play an essential role in looking at ways to reduce the redundancy in data collection and ensure that the data collected meet the needs of multiple users. NCES has worked with other units within the Department of Education, but now they should reach beyond the education boundaries. Data providers will greatly appreciate any assistance that NCES can provide in convening and urging agreement on data formats and in considering ways that data can be legally shared with health and human services.

NCES can serve the education community well by keeping a focus on the future and what must be done to ensure that data collection efforts take advantage of electronic advances and meet future information needs.

DENNIS CARROLL

Fritz Scheuren's paper describes several opportunities for NCES. He broadly and boldly develops major implications for operations, staffing, and technology. Whether his predictions are realized within the next 10 years or not, NCES should prepare for the next revolution in analysis. This revolution is not statistical technique, but rather the predominance of administrative records as the birthing agent for data sets.

The paper rightly suggests that the quantity of administrative record data that may be tapped by NCES will continue to increase. Further, with faster, cheaper, and better connected computing, administrative records will be easier to use. Scheuren suggests that eventually data collections may become supplements for administrative data rather than the currently reversed situation. However, Scheuren failed to note the impact of restrictive privacy legislation, state budget declines, reinvention, and other political factors that are increasingly restricting access to systems of administrative records.

If Scheuren's notions are attempted, NCES must consider how far on the leading edge of this technological adventure it should venture. With limited budgets, NCES needs the administrative data to enhance limited data collections. However, with a shrinking staff and an apolitical mission, it is difficult to meet the demands of leading-edge status. The paper would be improved if it included suggestions about the areas NCES should try initially.

With an increase in administrative record quantity, there will be a compatibility potential that is limited by comparability. Imputations, as suggested in the paper, will become more prevalent. Without significant advances in imputation technology, the notions of fully or partially imputed data sets will be limited. Currently, it is doubtful that a little reported data can be appropriately combined with a lot of imputed data for meaningful analyses. For example, although imputation makes a constructed NPSAS possible with Central Processing System and IPEDS data as a source, the policy community probably would not use it.

Just as instrument nonresponse plagues survey collections, partial access will trouble administrative records. Biases associated with instrument nonresponse rarely have the impact of restrictions on access to administrative records. Analysts with access hold an advantage over those using the biased, even if fully imputed, data. How NCES should deal with this conflict is an important issue.

Finally, this paper rightly suggests that getting distributions "correct" should be more important to NCES than cleaning data case by case and variable by variable. Well-behaved data that adequately reflect the proper distribution(s) are simply better. Error estimation, modeling, and simple statistics (graphical displays) feel better when using well-distributed data. In this area, administrative records can help, and they can help immediately. Many distributions can be known based on administrative records, without access to the microdata.

WILLIAM H. FREUND

Glynn Ligon was given the impossible task of describing "new developments in technology that have affected or will affect the collection and reporting of education data." This represented a difficult assignment at best and was impractical in this era of highly evolving telecommunications and eventual saturation of computers into our work and home environments. The issues are not technological changes—we know these will occur. Since these changes, particularly in telecommunications, will open up new markets for education statistics, the more important questions for NCES include the following:

- Who will be the customers of national education data?
- What questions will they ask?
- How should information be presented and retrieved?

It is important to note that these three questions do not even address the mechanics of technology (hardware and software). We will have the technology; the only issue is the extent of access within the education community and our customer base to this technology. Access is an important question for schools and districts without the financial resources to obtain high-speed Internet connections.

However, assuming access, just exactly how would these technologies affect the Center's data collection and dissemination of administrative records survey data? And is the Center doing anything now to take advantage of what is available?

Data collections for administrative records

Many people think that NCES continues to rely on paper forms for much of its data collection/survey work. Currently, the Center uses at least five different modes to obtain information from state agencies and colleges and universities. These include DBF files, ASCII-based data (on diskettes or tape), File Transfer Protocol (FTP), mail, and Electronic Data Interchange (EDI). However, only in library collections have we moved beyond these five somewhat traditional modes into an electronic forms mode. Only our library programs have turned in this direction, but plans are now under way to move more actively into electronic forms. At present, there are many "software" models available to guide our developmental efforts, specifically packages such as TurboTax™. These packages provide forms, year-to-year comparisons, and internal editing capabilities for consistency of responses.

But the important thing to remember is the impact of shifting to new collection practices. Technology will force data owners and providers to assume more responsibility for data quality and timeliness. Thus, NCES's responsibility will shift toward developing and providing data owners with new and better tools to improve quality and timeliness.

Dissemination of Administrative Records

As with data collection activities, there is a misperception about how NCES disseminates its products. Computer tapes are no longer our primary mode of dissemination. In fact, we prefer *not* to send tapes. However, we are awash in new forms of products, including diskettes, CD-ROMs, tabulation packages (the Data Analysis System), Electronic Codebooks (ECBs), printed reports, gopher servers, phone orders, and, yes, a few tapes. In fact, these new products are invaluable to our customers. For example, the DAS software developed by Dennis Carroll and Larry Bobbitt obviates the need for users to understand complex samples, since the software handles the appropriate calculations for variances.

New techniques or methods are coming. For example, we are developing a World Wide Web (WWW) home page. We are also setting up an early release program for administrative records. And we are improving customer service in other ways, including expanding of the National Data Resource Center (NDRC). The NDRC provides tabulation services to customers without access to computers and/or appropriate software packages. But our real future in dissemination is embodied by our current initiatives with Structured Query Language (SQL) server and data warehouses.

Envision sitting in front of your personal computer; loading Excel onto your desktop; clicking on external data; linking to NCES via Internet; selecting data files of your choice; subsetting the file based on your own criteria; tagging those data elements that you want; and then retrieving the data back into your Excel spreadsheet. That scenario will be the ultimate dissemination program—providing the user with the right information, in the right form, in the right place, and at the right time. That scenario is actually viable today and is being tested internally within NCES and externally via point-to-point protocol.

Glynn Ligon's paper hits home on a variety of issues before these scenarios become a practical reality. First, you must be very familiar with file structures to use SQL server—user friendliness is not a design feature when it comes to data. Second, the user must have excellent documentation to use the files effectively. Electronic codebooks and DAS CD-ROMs are a step in that direction. But we should convert them to Windows so that users will simply press the F1 help key to obtain full descriptions of variable definitions and values. Another issue is for NCES to fully understand its customer capabilities. We might, as suggested by Fritz Scheuren, use the Common Core of Data (CCD) and Integrated Postsecondary Education Data System (IPEDS) to periodically survey our respondents and customers. We would then have some answers to the questions raised at the beginning of this commentary.

But easier data collections and expanded user access to data raise additional areas for the Center to consider and act upon. For example, standards and data comparability among survey respondents will become increasingly important. This is true across all levels of education, and NCES is currently promoting comparability via its efforts with the Cooperatives, handbooks, and EDI standards. We also have to promote more leveraging of software if survey respondents are to make effective use of new technologies. While the cooperatives can play a role in this effort, responsibility will fall upon the states themselves. Finally, NCES must help users DIRTFT—Do It Right The First Time. In this case, "It" means drawing valid conclusions or findings from the various NCES data files.

With all these activities under way, NCES is addressing the challenges imposed by new technologies. I wonder what form those challenges will assume 5 years from now?

A Appendix A
About the Contributors

About the Contributors

AUTHORS

Robert F. Boruch is University Chair Professor in the Graduate School of Education and the Statistics Department of the Wharton School at the University of Pennsylvania. A Fellow of the American Statistical Association, he has received awards for his work on research methods and policy from the American Educational Research Association, the American Evaluation Association, and the Policy Studies Association. Boruch is the author of nearly 150 scholarly papers and the author or editor of a dozen books on topics ranging from evaluation of AIDS prevention programs and social experiments to assuring confidentiality of data in social research.

David W. Breneman is University Professor and Dean of the Curry School of Education at the University of Virginia. He was Visiting Professor at the Harvard Graduate School of Education from 1990 to 1995, where he taught graduate courses on the economics and financing of higher education, on liberal arts colleges, and on the college presidency. As a Visiting Fellow at the Brookings Institution, he conducted research for a book entitled *Liberal Arts Colleges: Thriving, Surviving, or Endangered?*, published by Brookings in 1994. From 1983 to 1989, Breneman served as president of Kalamazoo College, a liberal arts college in Michigan. Prior to that, he was a Senior Fellow at Brookings, specializing in the economics of higher education and public policy toward education. From 1972 to 1975, he was Staff Director of the National Board on Graduate Education at the National Academy of Sciences, where he wrote on policy issues confronting graduate education. In addition, Breneman served as Executive Editor of *Change,* the magazine of higher learning. Dr. Breneman received his bachelor's degree in Philosophy from the University of Colorado and his Ph.D. in Economics from the University of California at Berkeley, and taught at Amherst College before moving to Washington in 1972.

Dominic Brewer is a Labor Economist specializing in the economics of education and education finance. His research has focused on educational productivity and teacher incentives, using large national databases such as High School and Beyond and the National Education Longitudinal Study of 1988. Examples of this work include an analysis of the effects of teacher education and quality on student achievement gains, the interaction between student and teacher race, gender and ethnicity, the effects of ability grouping on student achievement, and the effects of administrative resources on student performance. He has published numerous articles in academic journals such as *Review of Economics and Statistics, Journal of Labor Economics, Industrial and Labor Relations Review*, and *Economics of Education Review*, as well as other publications such as *Phi Delta Kappan.* Dr. Brewer received a Ph.D. in Labor Economics from Cornell in 1994, and holds a bachelor's degree from Oxford University. He has been an Associate Economist at

RAND since 1994 and is also a Visiting Assistant Professor of Economics at the University of California, Los Angeles.

Peter Cappelli is Co-Director of the National Center on the Educational Quality of the Workforce (EQW) at the University of Pennsylvania.

Christopher T. Cross is President of the Council for Basic Education (CBE) as well as President of the Maryland State Board of Education. Before joining CBE, Mr. Cross served as Director of the Education Initiative for The Business Roundtable and as Assistant Secretary of Education Research and Improvement (OERI) in the U.S. Department of Education. At OERI, he was responsible for the research, statistical, and improvement programs of the Department of Education. He joined the federal government for the first time in 1969 with the U.S. Department of Health, Education, and Welfare, where he served as Deputy Assistant Secretary for Legislation. From 1973 to 1978, Mr. Cross served as the Senior Education Consultant and Republican Staff Director of the Committee on Education and Labor, U.S. House of Representatives. Mr. Cross has written extensively in the education and public policy areas, and his articles have appeared in numerous scholarly and technical publications. Mr. Cross earned a bachelor's degree from Whittier College and a master's degree in Government from California State University in Los Angeles.

Fred J. Galloway is the Director of Federal Policy Analysis at the American Council on Education (ACE). In this position, he represents the interests of the higher education community before the executive and legislative branches of the federal government and is responsible for analyzing the effects of legislation on colleges and universities. Before joining ACE, Dr. Galloway was a member of the faculty of the Economics Department at San Diego State University and of the School of Business at the University of San Diego. Dr. Galloway received both a bachelor's and master's degree from the University of California at San Diego, and a doctorate from Harvard University.

Gary Hoachlander is President of MPR Associates, Inc., a consulting firm specializing in management, planning, and research for a variety of public and private clients. A nationally known expert on vocational education and preparation for work, he also serves as MPR Associates' site director for the work performed by the firm for the National Center for Research in Vocational Education at the University of California at Berkeley. He has conducted research and published on a wide variety of issues including industry-based curriculum, industry skill standards, performance measures and assessment, finance, and national education data systems. Dr. Hoachlander received his Ph.D. in City and Regional Planning from the University of California, Berkeley. He also holds a master's degree in City Planning from U.C. Berkeley, and earned his bachelor's degree from Princeton University, where he attended the Woodrow Wilson School for Public and International Affairs.

John F. Jennings is the Director of the Center on National Education Policy. The Center's purpose is to inform the general public, educators, and policymakers of the developments in school reform across the country and also of the changes in federal education programs. From 1967 to 1994, Mr. Jennings worked in the area of federal aid to education for the U.S. Congress. In that capacity, he was involved for the last 25 years in nearly every major education debate held at the national level as well as the reauthorizations of the major federal education programs including the Elementary and Secondary Education Act, the Vocational Education Act, the School Lunch Act, the Individuals with Disabilities Education Act, and the Higher Education Act. Mr. Jennings has also edited several books, published numerous articles, and writes a national newsletter.

Glynn D. Ligon is President of Evaluation Software Publishing, Incorporated.

David R. Mandel is Vice President for Policy Development at the National Board for Professional Teaching Standards in Washington, D.C., where he has primary responsibility for overseeing the Board's standards development efforts and education policy and reform program. Previously, Mr. Mandel was Associate Director of the Carnegie Forum on Education and the Economy; a Senior Policy Analyst in the Office of the Under Secretary of Education; and the National Institute of Education's Assistant Director responsible for managing the Institute's research program in education finance, governance, and human capital. He began working on education policy issues in the early 70s at the U.S. Office of Economic Opportunity, where his efforts were directed at the needs of poor and minority children.

Charles E. Metcalf is President of Mathematica Policy Research, Inc., which is one of the nation's leading independent research firms and conducts public policy research and surveys for federal and state governments as well as private clients. He is nationally known for his research on social experimentation and income distribution and has directed research activities at Mathematica for the past 21 years. Dr. Metcalf specializes in experimental and sample design, data collection design, and analytic design efforts. His expertise, gleaned from 28 years of experience in the field, spans all facets of research design and analysis. He has played a major role in more than 30 major social experiments, demonstrations, and evaluations. Dr. Metcalf has a Ph.D. in Economics from the Massachusetts Institute of Technology.

Morton Owen Schapiro is Professor of Economics and Dean of the College of Letters, Arts, and Sciences at the University of Southern California. He and Michael McPherson have co-authored two recent books on American higher education: *Keeping College Affordable: Government and Educational Opportunity* (Brookings 1991), and (with Gordon Winston), *Paying the Piper: Productivity, Incentives, and Financing in U.S. Higher Education* (University of Michigan Press 1993).

Fritz Scheuren has extensive experience in using administrative records in sample surveys and other settings. Currently, Visiting Professor of Statistics at The George Washington University, Dr. Scheuren retired in 1994 as Director of the Statistics of the Income Division of the Internal Revenue Service. Formerly, he had been the Chief Mathematical Statistician at the Social Security Administration. In 1995, he won the Shiskin Award for contributions to U.S. economic statistics and among other honors is a Fellow of the American Statistical Association and the American Association for the Advancement of Science. He has published more than 90 papers, monographs, and books—both applied and theoretical—mainly in the area of survey sample design and estimation, including such topics as record linkage, privacy, and the handling of missing data. He holds a master's and doctoral degree in Statistics from The George Washington University.

Diane Stark is the Associate Director of the Center on National Education Policy. From 1988 to 1994, Ms. Stark was a legislative associate for the U.S. House of Representatives Committee on Education and Labor, where she assisted in the reauthorization of the major federal education programs. Prior to her work in the Congress, she was employed in the government relations offices of the National PTA and the Council of Chief State School Officers.

Cathleen Stasz is a Senior Behavioral Scientist at RAND and Site Director for the National Center for Research in Vocational Education (NCRVE). Her research areas include the implementation of advanced computer-based technologies in education, the workplace and the military, systemic school reform, and teaching and learning generic skills for the workplace. Currently, her projects include a study of the determinants of employer participation in school-to-work programs and an examination of the quality of student experiences in work-based learning environments.

Amy Rukea Stempel, Assistant Director for Standards Analysis at the Council for Basic Education (CBE), has been affiliated with the Council since 1989. In 1992, she left the CBE to teach the International Baccalaureate (English literature) at the Kodaikanal International School, Kodaikanal, India, and then returned to the CBE in the fall of 1994. Ms. Stempel has published numerous articles that inform the academic standards-setting process and the relationship of various education reforms to academic learning in CBE's flagship publication *Basic Education* and in *Teacher Magazine*. In addition, she designed the popular CBE chart "Standards: A Vision for Learning" (spring 1991), which synthesized all the current standards projects and was reprinted in 1994. A candidate for a master's degree in the Humanities at Georgetown University, Ms. Stempel is primarily engaged in writing about education reform and managing standards projects at CBE. She has a bachelor's degree in English from Carnegie Mellon University.

James W. Stigler is Professor of psychology at UCLA and Director of the Third International Mathematics and Science Study (TIMMSS) Videotape Classroom Study.

George Terhanian, a doctoral candidate at the University of Pennsylvania, is presently serving as an American Education Research Association Research Fellow at the National Center for Education Statistics. His general research interest lies in synthesizing evidence generated by local experiments and nationally representative surveys. Mr Terhanian has several years of teaching and administrative experience in public and private schools.

DISCUSSANTS

Sharon Bobbitt received her Doctorate in Education Research from the University of Virginia in 1986. She worked for 8 years on the Schools and Staffing Survey, with a primary focus on teacher issues. She is currently Director of the Knowledge Applications Division in the Office of Educational Research and Improvement.

Dennis Carroll earned a Ph.D. in mathematics and quantitative psychology from Vanderbilt University in 1974. Since 1980, he has managed the National Center for Education Statistics' longitudinal studies program where several projects have incorporated records data.

Barbara S. Clements is Acting Director of the State Education Assessment Center at the Council of Chief State School Officers (CCSSO). She also directs the National Elementary/Secondary Education Data and Information System Project, which is funded by the U.S. Department of Education's National Center for Education Statistics to promote the standardization, automation, and effective utilization of data about education. Before joining the CCSSO staff, Dr. Clements worked on the development and administration of teacher assessment and evaluation instruments for the state education agency in Texas. She is a co-author of two textbooks on effective classroom management, soon to be released in their fourth edition. Dr. Clements holds a bachelor's degree in Education from the University of Texas at Austin, and is certified to teach secondary Spanish and Government. In addition, she has master's degree in Foreign Language Education and a Ph.D. in Educational Psychology from the University of Texas at Austin.

Emerson J. Elliott is a consultant on education policy, Federal statistics and management. He left the Federal Government in 1995 after a career that included heading the National Center for Education Statistics nearly eleven years and serving as the first "Commissioner of Education Statistics" when that post became a Presidentially appointed, Senate confirmed position under legislation enacted in 1988. Previously he had led the Issues Analysis Staff in the Office of the Under Secretary of Education, served as the Deputy Director of the National Institute of Education, and directed the OMB education branch when that was established in 1967.

Mary Frase is the Senior Technical Advisor in the Data Development and Longitudinal Studies Group, National Center for Education Statistics, in the U.S. Department of Education. Prior to joining NCES in 1985, she was a faculty member at Teachers College, Columbia University, and worked as an independent consultant advising state and local governments and conducting research in the areas of education policy, education finance, and state-local finance.

William H. Freund works within the U.S. Department of Education's National Center for Education Statistics. He recently assumed responsibility for adapting information technologies into the Center's data collections, program administration, and information dissemination. Just before this new position, he was responsible for institutional studies of postsecondary education. In that capacity, he was the program manager for the Integrated Postsecondary Education Data System (IPEDS)—a series of annual statistical surveys that collect enrollment, completions, finance, salary, and staffing data from the nation's postsecondary education institutions.

Paula R. Knepper is a Statistician in the Postsecondary Longitudinal Studies department of the National Center for Education Statistics.

James F. McKenney is currently the Director of Workforce Development, formerly the Office of College Employer Relations, at the American Association of Community Colleges (AACC). Also, he has served as the Assistant Vice President for Federal Relations, with responsibilities for the reauthorization of the Carl Perkins Vocational Education Act and the Job Training Partnership Act. As Director of Workforce Development, Dr. McKenney is charged with being the primary liaison between AACC and the various relevant federal departments and trade associations. In this role, Dr. McKenney has continued to track the implementation of the various federal human resource development laws. He received his bachelor's and master's degrees from the University of Florida and his doctorate from the University of Maryland.

Michael McPherson is the Dean of the faculty at Williams College. He is W. van Alan Clark Third Century Professor of Economics and Co-Director of the Williams Project on the Economics of Higher Education. Earlier, he served as Chair of the Williams Economics Department, as Senior Fellow in Economic Studies at the Brookings Institution and as Fellow of the Institute for Advanced Study. Mr. McPherson is co-author of two recent books, *Keeping College Affordable: Government and Educational Opportunity* (Brookings 1991) and *Paying the Piper: Productivity, Incentives and Financing in American Higher Education* (University of Michigan Press 1993). His new book, *Economic Analysis and Moral Philosophy*, co-authored with Daniel Hausman, was published by Cambridge University Press in 1996.

Jamie P. Merisotis is the founding President of the Institute for Higher Education Policy in Washington, D.C. The Institute is a non-profit, non-partisan organization with the mission of fostering access to and quality in postsecondary education through the development and promotion of innovative solutions to the important and complex issues facing higher education. The Institute has conducted a number of recent studies including *The Next Step: Student Aid for Student Success; College Debt and the American Family; Enhancing Quality in Higher Education;* and *Affirmative Action and the Distribution of Resources in U.S. Department of Education Programs.*

Kevin Miller is currently Associate Professor of Psychology at the Beckman Institute at the University of Illinois, Urbana-Champaign. His research interests concern the effects of symbolic tools on cognitive development, focusing on how language and cultural differences between China and the United States affect the development of abilities such as reading and mathematical competence. He received his Ph.D. from the University of Minnesota, and then taught at Michigan State University and the University of Texas at Austin before joining the faculty at the University of Illinois. His research is currently supported by a Research Scientist Development Award and a research grant, both from the National Institute of Mental Health.

Frederick Mosteller is Roger I. Lee Professor of Mathematical Statistics Emeritus, Harvard University. He directs the Center for Evaluation of the Initiatives for Children Project at the American Academy of Arts and Sciences. Over the years, his research work has been devoted to theoretical and applied statistics. Dr. Mosteller works in data analysis, meta-analysis, robust methods, health and medicine, and social sciences, and has also written on sports statistics. While at Harvard, he has chaired the departments of Statistics, Biostatistics, and Health Policy and Management.

Mary Rollefson is a senior survey analyst with the National Center for Education Statistics. She has published several reports on teacher supply and demand and serves as the NCES liaison to the National Education Goals Panel.

Donald B. Rubin is Professor in the Department of Statistics, Harvard University. He has written nearly 250 publications (including several books) on a variety of topics, including computational methods, causal inference, survey methods, techniques for handling missing data, Bayesian methods, multiple imputation, matched sampling, and applications in many areas of social and biomedical science. Professor Rubin is a Fellow of the American Statistical Association, the Institute for Mathematical Statistics, the International Statistical Institute, the Woodrow Wilson Society, the John Simon Guggenheim Society, the New York Academy of Sciences, the American Association for the Advancement of Sciences, and the American Academy of Arts and Sciences. He is also the recipient of two of the most prestigious awards available to statisticians: the Samuel S. Wilks Medal of the American Statistical Association and the Parzen Prize for Statistical Innovation.

Eileen Mary Sclan is currently an Assistant Professor of Education in the Department of Curriculum and Instruction at Long Island University—C.W. Post Campus. Her main areas of research interest include teachers' workplace conditions, teacher performance evaluation, and teacher induction. At present, she is analyzing national data (funded by an AERA/NCES grant) to examine the inequitable distribution of qualified teachers and workplace supports. Dr. Sclan received her Ed.D. in Educational Leadership from Teachers College, Columbia University.

David Stern is Professor of Education at the University of California at Berkeley, and Director of the National Center for Research in Vocational Education, based at Berkeley's Graduate School of Education. From 1993 to 1995, he was principal administrator in the Center for Educational Research and Innovation at the Organization for Economic Cooperation and

Development in Paris. Since 1976, he has been on the faculty at Berkeley, teaching and conducting research on the relationship between education and work, and on resource allocation in schools. David Stern is the lead author of several recent books: *School to Work: Research Programs in the United States* (with N. Finkelstein, J. Stone III, J. Latting, and C. Dornsife 1995); *School-Based Enterprise: Productive Learning in American High Schools* (with J. Stone III, C. Hopkins, M. McMillion, and R. Crain 1994); and *Career Academies: Partnerships for Reconstructing American High Schools* (with M. Raby and C. Dayton 1992). He also co-edited *Market Failure in Training* (with J.M.M. Rtizen 1991), and *Adolescence and Work: Influences of Social Structure, Labor Markets, and Culture* (with D. Eichorn 1989).

P. Michael Timpane, Vice President of the Carnegie Foundation, is involved in developing all aspects of the programs of the Foundation. In his own research, he is assessing the progress and problems of contemporary national education reform. Mr. Timpane is also Professor of Education and former President of Teachers College, Columbia University, the world's most comprehensive graduate school for the preparation of educational, psychological, and health professionals. Previously, he served as Dean of Teachers College and as Deputy Director and Director of the federal government's National Institute of Education. He has conducted research on educational policy as a senior staff member at the Brookings Institution and the RAND Corporation. Also, Mr. Timpane is a member of the Pew Forum on Education Reform, for which he is currently organizing and editing a volume of essays on higher education's involvement in precollegiate school reform. In addition, he serves on the boards of Children's Television Workshop, the Southern Education Foundation, the Synergos Institute, and Jobs for Education and the American Associate of Higher Education. Mr. Timpane received a bachelor's and a master's degree in history from Catholic University, and an M.P.A. degree from Harvard University in 1970. He has received honorary doctorates from Wagner College and Catholic University.

B

Appendix B
Future NCES Data Collection Conference Agenda

National Center for Education Statistics

Future NCES Data Collection: Some Possible Directions
Conference Agenda, November 27–29, 1995

Hyatt Regency Washington on Capitol Hill
400 New Jersey Avenue, N.W., Washington, DC 20001

Monday, November 27
7:00 p.m. Dinner in the Ticonderoga Room
Emerson Elliott will address the group after dinner.

Tuesday, November 28
8:30 a.m. Continental breakfast. Congressional B meeting room
Welcome and opening remarks by Jeanne Griffith.

9:00–10:30 a.m.
Session 1—Tracking Education Reform: Implications for Collecting National Data Through 2010
First Paper: Jack Jennings and Diane Stark
Second Paper: Chris Cross and Amy Stempel
External Discussant: Tom Kane
Internal Discussant: Mary Frase

10:30–10:45 a.m.—Break

10:45 a.m.–12:30 p.m.
Session 2—Curriculum, Pedagogy, and Professional Development
First Paper: Curriculum and Pedagogy: Implications for National Surveys
Authors: Cathy Stasz and Dominic Brewer
Second Paper: Teacher Education, Training, and Staff Development: Implications for National Surveys
Author: David Mandel
External Discussants: Michael Timpane and Eileen Sclan
Internal Discussants: Mary Rollefson and Sharon Bobbitt

12:30–1:30 p.m.—Lunch

1:30–2:45 p.m.
Session 3—Trends in Statistical and Analytic Methodology: Implications for National Surveys
Authors: Bob Boruch, George Terhanian, and Others

External Discussant: Fred Mosteller
Internal Discussant: Sue Ahmed

2:45–3:00 p.m.—Break

3:00–4:15 p.m.
Session 4—New Data Collection Methodologies, Part II: Experimental Design
Author: Chuck Metcalf
External Discussant: Don Rubin
Internal Discussants: Joe Conaty and Marilyn McMillen

4:15–4:30 p.m. Wrap-up first day

Wednesday, November 29—Congressional B meeting room

8:30–8:45 a.m.—Continental Breakfast

8:45–10:30 a.m.
Session 5—Postsecondary Education
First Paper: Tracking the Costs and Benefits of Postsecondary Education: Implications for National Surveys
Authors: Michael McPherson and Morty Schapiro
Second Paper: Special Issues in Postsecondary Education and Lifelong Learning: Implications for National Surveys
Authors: David Breneman and Fred Galloway
External Discussants: Jamie Merisotis and Jim McKenney
Internal Discussants: Roz Korb and Paula Knepper

10:30–10:45 a.m.—Break

10:45 a.m.–12:00 p.m.
Session 6—New Data Collection Methodologies, Part I: Observational Strategies
Author: Jim Stigler
External Discussant: Kevin Miller
Internal Discussant: Lois Peak

12:00–1:00 p.m.—Lunch

1:00–2:00 p.m.
Session 7—Education for Work: Curriculum, Performance, and Labor Market Outcomes
Author: Peter Cappelli
External Discussant: David Stern
Internal Discussants: Nabeel Alsalam and Marilyn Binkley

2:00–2:15 p.m.—Break

2:15–3:45 p.m.

Session 8—Using Administrative Records and New Developments in Technology

First Paper: Opportunities for Making More Effects Use of Administrative Records in Surveys of Elementary, Secondary, and Postsecondary Education

Author: Fritz Scheuren

Second Paper: New Developments in Technology: Implications for Collecting, Storing, Retrieving, and Disseminating National Data for Education

Author: Glynn Ligon

External Discussant: Barbara Clements

Internal Discussants: Dennis Carroll and Bill Freud

3:45–4:00 p.m.—Conference Wrap-up

☆ U.S. GOVERNMENT PRINTING OFFICE: 1996 – 413 - 470 / 60009

United States
Department of Education
Washington, DC 20208–5650

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G–17

Special Fourth Class

BEST COPY AVAILABLE